

The University of San Francisco

META-ANALYSIS OF THE EFFECTIVENESS  
OF TASK-BASED INTERACTION  
IN FORM-FOCUSED INSTRUCTION OF ADULT LEARNERS  
IN FOREIGN AND SECOND LANGUAGE TEACHING

A Dissertation Presented  
to  
The Faculty of the School of Education  
Learning and Instruction Department

In Partial Fulfillment  
of the Requirements for the Degree  
Doctor of Education

by  
Marina Cobb  
San Francisco  
December 2010

## ABSTRACT

Research into the effectiveness of task-based interaction in acquisition of specific grammatical structures of the target language has been scarce and sometimes has presented conflicting findings. Task-based interaction engages learners in focused face-to-face oral-communication tasks that predispose them to repeated use of the target structure in meaningful contexts.

Previous meta-analyses have provided some evidence of effectiveness of task-based interaction in learners' morphosyntactic development (Keck, Iberri-Shea, Tracy-Ventura, & Wa-Mbaleka, 2006; Mackey & Goo, 2007). The present meta-analysis adopts a somewhat different perspective from one or both of the previous meta-analyses through the following features: exclusion of studies that focus only on effects of corrective feedback, inclusion of both published and unpublished studies to expand the search domain, imposing of more stringent criteria for oral-communication tasks, focusing on adult learners and face-to-face, rather than computer-mediated interaction, and so forth.

This meta-analysis synthesized the results of 15 primary studies. On average, learners who received task-based interaction treatments through completing focused oral-communication tasks with native or nonnative interlocutors performed better than learners who received no focused instruction in the target structure and somewhat better than learners who received other types of instruction such as traditional grammar instruction, input processing activities, and so forth. The effect sizes were medium and small, respectively. Both the learners who received task-based interaction and those who received other instruction showed large within-group

gains, whereas the gains demonstrated by the learners who received no instruction in the targeted form were insignificant or small based on Cohen's 1977 classification.

The effects of task-based instruction were durable.

The analysis of the characteristics of tasks, target structures, educational settings, and so forth as moderator variables has identified statistically significant differences for some of these factors. The analog to the analysis of variance identified the complexity of the target structure, the nature of participant assignment to groups (nonrandom vs. random), and the difference between long-delay and short-delay posttests as factors that can account for variability in effect sizes. The meta-analytic findings expanded the scope of understanding of the effects of task-based interaction and were instrumental in formulating suggestions for future research in the domain.

The dissertation, written under the direction of the candidate's dissertation committee and approved by the members of the committee, has been presented to and accepted by the Faculty of the school of Education in partial fulfillment of the requirements for the degree of Doctor of Education. The content and research methodologies presented in this work represent the work of the candidate alone.

Marina Cobb

Candidate

November 2, 2010

Date

Dissertation Committee

Dr. Patricia Busk

Chairperson

November 2, 2010

Dr. Lanna Andrews

November 2, 2010

Dr. Stephen Cary

November 2, 2010

DEDICATION

To my parents

Ludmila Evgenievna Nikolaeva

and

Lev Konstantinovich Nikolaev

ПОСВЯЩАЕТСЯ

моим родителям

Людмиле Евгеньевне Николаевой

и

Льву Константиновичу Николаеву

## ACKNOWLEDGEMENTS

There are many people to whom I wish to express my gratitude. First, I would like to thank Dr. Patricia Busk, the Chair of my Dissertation Committee, who also was the teacher who introduced me to meta-analysis. I am thankful for all her expertise, support, patience, and kindness.

Many thanks to my Dissertation Committee Members, Dr. Lanna Andrews and Dr. Stephen Cary, for supporting my research interest enthusiastically and for their insightful comments. Their encouragement meant a lot to me.

I am very grateful to the Defense Language Institute Foreign Language Center (DLIFLC) where I work for supporting my tuition through its tuition-assistance program. I simply would not have been able to enroll in the doctoral program at the University of San Francisco without this assistance. I also am grateful to all the supervisors and managers at DLIFLC who have supported my academic endeavors.

I am honored to have been able to receive assistance and encouragement from renowned experts in the field of second language acquisition Dr. Patsy Lightbown and Dr. Nina Spada, as well as Yasuyo Tomita, who was Dr. Spada's doctoral student at the time. I am grateful to all primary research study authors who answered my requests for dissertation copies or additional information regardless of whether my inclusion criteria made it possible for me to include their studies in the meta-analysis in the end. It was uplifting to encounter such responsiveness and support from the academic community.

Among the people whom I have met through this process one name, in particular, stands out. I wish to extend a very special thank you to Luke Plonsky, a doctoral student at Michigan State University, who has provided truly invaluable assistance by answering

my questions about meta-analytic procedures as they relate to the field of second language acquisition. I was very impressed by his expertise, responsiveness, and willingness to help. Thank you to Dr. Nicole Tracy-Ventura for introducing me to him.

I wish to express my gratitude to Natalie Lovick, my colleague and friend who, without hesitation, found time in her very busy schedule to serve as the second coder in my meta-analysis. Her hard work and insightful suggestions were very much appreciated.

There was one more very special person who has made a significant contribution. A big thank you for invaluable assistance goes to my son Constantine Perepelitsa who is a UC Santa Cruz computer-science student and a research assistant at the Naval Postgraduate School. The amazing fact about Constantine's involvement in my dissertation project is that he did not only write code for my calculations and check my Excel formulas, which, of course, is in line with his expertise. He also helped input or transfer data associated with thousands of lines of calculations during the difficult times brought about by my injury. Moreover, he provided editorial and proofreading assistance, as usual, leaving me in awe of his ability to comprehend material in a field (second language acquisition) in which he does not have expertise. His great sense of humor was greatly appreciated as well, and I can confirm that he truly deserves the "King of Sarcasm" title lovingly awarded him by the graduate students at the Naval Postgraduate School.

To my other son Dennis Perepelitsa, who himself is a doctoral student at Columbia University and my favorite source of knowledge of all things: yes, I technically "beat" you to a doctoral degree in terms of time but only because I had had a head start. Otherwise, I never stood a chance.

I am grateful to my daughters, Christina and Polina, for their graciousness, understanding, and boundless patience. Thank you for loving school and never letting me feel alone while doing my homework. Additionally, I am thankful to all my friends and colleagues for encouragement, patience, and enthusiasm about my research.

To my giving husband Chris, there are simply no words to describe your enormous contribution. Thank you for always being there.

Thank you to my parents whose voices beam with pride during overseas calls. This dissertation is dedicated to you because you were there first. You think you taught me to be a good student through what you said but it was really through what I saw you do and through who you are.

While I am grateful to the people named here for their contributions to this research project, all errors and omissions, of course, are my sole responsibility.



## TABLE OF CONTENTS

	Page
ABSTRACT .....	ii
DEDICATION.....	v
ACKNOWLEDGEMENTS .....	vi
TABLE OF CONTENTS.....	ix
LIST OF TABLES.....	xiii
LIST OF FIGURES.....	xv
CHAPTER	
I. RESEARCH PROBLEM .....	1
Statement of the Problem .....	2
Purpose of the Study.....	8
Theoretical Rationale .....	14
Task-Based Language Teaching .....	15
Focus on Form .....	16
Background and Need .....	18
Norris and Ortega’s Research Synthesis and Meta-Analysis of Effectiveness of L2 Instruction .....	19
Keck, Iberri-Shea, Tracy-Ventura, and Wa-Mbaleka’s Meta- Analysis Investigating the Empirical Link Between Task-Based Interaction and Acquisition.....	22
Mackey and Goo’s Research Review and Meta-Analysis of Interaction Research.....	24
Limitations of the Three Previous Meta-Analyses .....	27
Research Questions .....	34
Significance of the Study .....	35
Definition of Terms .....	38
Summary .....	45
Forecast of the Study .....	46
II. REVIEW OF LITERATURE .....	48
Historical Perspectives .....	50
Communicative Competence and Communicative Language Teaching (CLT).....	56
Role of Input and Output in Foreign and Second Language Learning.....	57
Role of Interaction in Foreign and Second Language Learning .....	63
Skill Acquisition in Foreign and Second Language .....	68
Task-Based Language Teaching (TBLT).....	75
Definition of Task .....	79

## TABLE OF CONTENTS Continued

CHAPTER	Page
Criterial Features of Tasks .....	81
Benefits and Limitations of TBLT .....	86
Types of Tasks as Moderator Variables .....	91
The Gap Principle and Major Task Designs .....	92
One-Way and Two-Way Tasks.....	96
Closed and Open Tasks.....	98
Divergent and Convergent Tasks.....	100
Focused and Nonfocused Tasks.....	101
Role of Individual Learner Differences in Task Performance .....	104
Other Task-Related Moderator Variables .....	107
General Considerations .....	108
Learner-to-Learner versus Teacher-led Interaction .....	109
Cognitive Complexity of the Task.....	111
Pedagogical Grammar and Language Acquisition .....	114
Focus on Forms, Focus on Form, and Focus on Meaning .....	115
Focus on Forms.....	115
Focus on Meaning.....	117
Focus on Form.....	119
Task-based Interaction as a Focus-on-Form Instructional Technique.....	121
Types of Target Structure as Moderator Variables.....	124
Degree of Task Essentialness of the Target Structure .....	130
Measures of Acquisition of Target Grammatical Structures.....	134
Common Data-collection Techniques in Outcome Measures ..	137
Naturalistic versus Elicited Data-collection Procedures..	137
Elicitation of Production Data.....	137
Elicitation of Comprehension Data.....	139
Elicitation of Metalinguistic Data.....	140
Types of Outcome Measures as Moderator Variables .....	141
Issues in Measuring Acquisition of Grammatical Structures ...	142
Review of Keck, Iberri-Shea, Tracy-Ventura, and Wa-Mbaleka's (2006) Meta-Analysis: Investigating the Empirical Link Between Task-Based Interaction and Acquisition .....	149
Summary .....	160
III. METHODOLOGY .....	163
Research Design .....	163
Data Sources and Search Strategies .....	168
Fail-Safe N .....	170
Inclusion and Exclusion Criteria .....	171
Inclusion Criteria .....	171

## TABLE OF CONTENTS Continued

CHAPTER	Page
Exclusion Criteria .....	172
Coding .....	173
Study Identification Information.....	174
Characteristics of the Outcome Measure .....	174
Methodological Features.....	177
Learner Characteristics .....	179
Treatment Design and Pedagogical Features .....	179
Quality of Study .....	180
Validity and Reliability of the Meta-Analysis .....	180
Validity .....	181
Interrater Reliability.....	183
Pretesting of the Coding Form .....	183
Data Analysis .....	184
Effect-Size Measures .....	184
Nonhomogeneity of Effect Sizes .....	189
Moderator Variables .....	190
Qualifications of the Researcher .....	191
IV. RESULTS.....	194
Research Synthesis .....	198
Research Publication.....	199
Research Setting and Context .....	200
Learner Characteristics .....	202
Methodological Features.....	204
Outcome Measures.....	209
Treatment Design and Pedagogical Features .....	214
Quantitative Meta-Analytic Findings .....	219
Calculating Independent Effect Sizes .....	220
Standardized-Mean-Difference Effect Size.....	221
Standardized-Mean-Gain Effect Size.....	228
Test of Homogeneity.....	232
Effects of Moderator Variables.....	236
Effects of Task Type.....	236
Task Type Based on the Gap Principle.....	236
Open-endedness and Convergence.....	239
Effects of Characteristics of Target Structures.....	241
Effects of the Duration of Treatment.....	243
Effects of Other Variables.....	244
Effects of Type of Outcome Measure.....	249
Summary .....	252

## TABLE OF CONTENTS Continued

CHAPTER	Page
V. DISCUSSION, LIMITATIONS, IMPLICATIONS, AND RECOMMENDATIONS .....	255
Summary of the Meta-Analysis.....	255
Limitations of the Study .....	257
Inclusion Criteria and Search Procedures .....	257
Small Number of Included Studies .....	260
Nonindependence of Study Samples and Effect Sizes .....	261
Disparity of Primary Study Designs .....	263
Methodological Quality of Included Studies .....	264
High-Inference Coding Decisions.....	265
Measurement Issues .....	267
Missing Data for Moderator Variables .....	270
Upward Bias for Standardized-Mean-Gain Effect Size.....	272
Discussion of Findings .....	274
Research Question 1 .....	274
Research Question 2 .....	279
Research Question 3 .....	281
Research Question 4 .....	284
Research Question 5 .....	288
Implications of the Study .....	291
Recommendations for Research .....	300
Conclusion.....	305
REFERENCES.....	307
APPENDIXES .....	333
APPENDIX A: Abbreviations .....	334
APPENDIX B: Additional Definitions of Terms .....	337
APPENDIX C: Coding Form.....	348
APPENDIX D: Draft Electronic Message Requesting a Copy of Study Report.....	368

## LIST OF TABLES

Table	Page
1. Relevant Second Language Acquisition (SLA) Hypotheses.....	67
2. Summary of Task Characteristics.....	105
3. Summary of Variables Potentially Affecting Learner Acquisition of the Target Structure Through Task-Based Interaction.....	134
4. Summary of Types of Outcome Measures.....	149
5. Overview of 15 Studies Included in the Present Meta-Analysis .....	195
6. Research Context, Target Language (TL), and Language Setting in Included Primary Studies.....	201
7. Study Design and Number of Participants in Included Studies.....	206
8. Types of Outcome Measures Used in Included Studies .....	213
9. Standardized-Mean-Difference Effect Sizes Calculated Based on the Contrasts Between Experimental and Control Groups .....	223
10. Standardized-Mean-Difference Effect Sizes Calculated Based on the Contrasts Between Experimental and Comparison Groups.....	226
11. Standardized-Mean-Difference Effect Sizes Calculated Based on the Contrasts Between Comparison and Control Groups.....	227
12. Standardized-Mean-Gain Effect Sizes Calculated for Experimental Groups.....	230
13. Standardized-Mean-Gain Effect Sizes Calculated for Control and Comparison Groups.....	233
14. Results of the Homogeneity Test ( <i>Q</i> Statistic) .....	234
15. Weighted Mean Effect Sizes for the Variable of Task Type (Based on the Gap Principle) and One-way versus Two-way Tasks.....	238
16. Weighted Mean Effect Sizes for the Variables of Open-endedness and Convergence .....	240

## LIST OF TABLES Continued

Table		Page
17.	Weighted Mean Effect Sizes Associated with Characteristics of the Target Structures .....	242
18.	Weighted Mean Effect Sizes Associated with the Duration of Task-Based Interaction Treatment.....	245
19.	Weighted Mean Effect Sizes Associated with Publication Type, Target Language (TL) and Language Setting, Research Setting, and Other Study-Related Variables .....	246
20.	Weighted Mean Effect Sizes Associated with Specific Types of Outcome Measures.....	250

## LIST OF FIGURES

Figure		Page
1.	Number of included studies by year of publication.....	200
2.	Frequency count for target languages (TLs) in included primary studies .....	202
3.	Box plot of standardized-mean-difference effect sizes.....	224
4.	Box plot of standardized-mean-gain effect sizes.....	231

## CHAPTER I

### RESEARCH PROBLEM

One of the challenges facing teachers of foreign and second languages is finding appropriate formats for teaching target language (TL) grammar within the current communicative methodology. The place of grammar in communicative language teaching (CLT; see Appendix A for a list of relevant abbreviations) frequently gives rise to differing positions and heated debates (Basturkmen, Loewen, & Ellis, 2004; DeKeyser, 2005; Doughty & Williams, 1998; R. Ellis, 2006a; Krashen, 1993; Lightbown, 2000; Swan, 2005).

Although teaching of grammar through interaction appears to be an accepted practice among some language teachers who received training in the West, others may hold more traditional beliefs about teaching grammatical language features (Hinkel & Fotos, 2002). Adherence to traditional methods is very strong in some parts of the world where teachers and students alike may equate learning grammar exclusively with discussions of intricate rules governing the language structure and with dissecting sentences and word forms. In particular, teachers of languages characterized by greater distance from the English language such as those that belong to Group III (e.g., Russian, Turkish, Persian-Farsi), Group IV (e.g., Arabic, Chinese, Korean), or Group V (e.g., Georgian) on a scale from I to V (MacWhinney, 1995) may be especially prone to such teaching beliefs.

For example, teachers of Russian traditionally attach a great value to explicit formal grammar instruction and, in particular, to teaching the contrastive analysis between the students' native language and the Russian language (Krouglov & Kurylko,



1999; Rodimkina, 1999). Their major argument appears to be that because Russian is an inflecting-fusional language in which grammatical endings simultaneously mark several grammatical categories (e.g., gender, number, case, animacy; Kempe & Brooks, 2008), Russian grammar does not lend itself well to communicative teaching. For this reason, teachers of Russian as well as other Group III to V languages may show resistance to implementing other, more communicative, instructional techniques for teaching grammar, thus neglecting the basic tenet of CLT that language should be taught through meaningful interaction as much as possible (Brandl, 2008; Canale & Swain, 1980; Savignon, 1983; Widdowson, 1978).

Kumaravadivelu (1994, 2003) who conceptualized so-called postmethod language teaching pedagogy (i.e., lack of adherence to one single methodology or recipe for instruction) for the era of communicative teaching underscored the need for situated pragmatism and principled eclecticism in the choice of classroom techniques. Nevertheless, the instructional practice of teaching grammar through interaction should occupy a fairly central role among other teaching practices aimed at developing the learners' grammatical competence (Doughty & Williams, 1998; Larsen-Freeman, 2001b; Lightbown, 2007; Long, 1996; Nassaji, 1999; Nassaji & Fotos, 2004; Spada, 1997). It is unrealistic, however, to expect language teachers to adopt this technique without solid empirical evidence of its effectiveness.

#### Statement of the Problem

Many adult students of foreign language (FL) and second language (L2) spend years learning the formal aspects of the TL (i.e., its phonetic system, verb conjugations, syntactical structure, etc.) in the classroom without ever developing an ability to function

in the TL, that is, to solve real-life problems, express ideas and feelings, or develop relationships with TL speakers (Adair-Hauck & Donato, 2002; Larsen-Freeman, 2003; Spada & Lightbown, 2008a). An equally unfortunate situation develops in the case of informal, or so-called street learners, immersed in the TL who develop a certain degree of fluency in the absence of grammatical accuracy. Although such learners may be able to satisfy some basic communicative needs, their ability to express more complex thoughts and to continue in their TL development is limited severely (Han, 2004; Higgs & Clifford, 1982). Therefore, the challenge is for FL and L2 teachers to find ways of developing the required grammatical accuracy and the ability to communicate at the same time, without sacrificing one or the other.

The traditional view on what constitutes grammar instruction is that no teaching takes place unless the teacher and the students engage in discussions of grammar rules, completing fill-in-the-blanks and other drills, or explicit analysis of sentence structure (Celce-Murcia, 1992; Larsen-Freeman, 2001a). Firmly opposed to this obsolete view, Widdowson (1988) asserted that properly conceived CLT does not neglect the teaching of grammar but rather recognizes its central mediating role in conveying meaning. Therefore, teaching of grammar should not be separated from meaningful classroom interaction. On the contrary, grammar should be taught through communication as much as possible (Adair-Hauck & Donato, 2002; R. Ellis, 2001, 2002; Larsen-Freeman, 2001a, 2001b, 2003; Nobuyoshi & Ellis, 1993). Research findings overwhelmingly support the assertion that, for the development of communicative ability, learners benefit from integration of form-focused activities with meaning-focused experiences, not exclusively

from one or the other (i.e., form-focused or meaning-focused), and not from one followed by the other (i.e., form-focused, then meaning-focused; Savignon, 2001).

Some recent research on teaching grammar specifically advocates the use of task-based learner interaction, that is, interactive form-focused activities that require the learners to produce output in the TL in pairs or small groups (Doughty, 2001; R. Ellis, 2001, 2002; Keck, Iberri-Shea, Tracy-Ventura, & Wa-Mbaleka, 2006; Kowal & Swain, 1994; Larsen-Freeman, 2003; Swain & Lapkin, 2001). These activities are *tasks*, rather than *exercises*, because they require learners to manipulate real-world information (vs. merely language form) while the learners communicate for a nonlinguistic goal in order to arrive at a nonlinguistic real-world outcome (Cobb & Lovick, 2007; R. Ellis, 2003; Nunan, 1989). These collaborative tasks frequently are referred to as *focused* tasks (R. Ellis, 2003; Nassaji & Fotos, 2004; Nobuyoshi & Ellis, 1993) because they are designed in such a way as to predispose learners to using the targeted language structure repeatedly.

Researchers also have referred to them as *focused communicative tasks* (R. Ellis, 2002) or *focused communication tasks* (Nobuyoshi & Ellis, 1993) to stress the point that task participants engage in meaningful communication with each other in the TL during task completion. Other researchers have used the term *structure-based* (tasks or activities) instead of *focused* or *form-focused*. For example, R. Ellis (2003) sometimes has referred to these activities as *structure-based production tasks*, Loschky and Bley-Vroman (1993) as *structure-based communication tasks*, and Fotos (2002) as *structure-based interactive tasks*. Other researchers emphasized the fact that these activities involve learner-produced *output* in the TL. For example, Koyanagi (1998) used the term *focus-*

*on-form output processing tasks*. Some of the so-called *structured output activities* as defined and advocated by Lee and VanPatten (2003) fall under the category of interactive form-focused tasks as well. Finally, Larsen-Freeman (2001a, 2003) coined a special term *grammaring tasks* to refer to such activities. Although there may be some differences in the definitions of these terms, for the most part, they are very similar and refer to classroom activities that *by design* (Nassaji, 1999) target improvement of the learners' structural accuracy in the TL and develop their ability to communicate meaning in the TL at the same time. In this study, the term *focused oral-communication tasks* was used. The role of *output* in second language acquisition (SLA), the concept of a language *task* (vs. *exercise*), and the definition of *focused* tasks as well as other related topics are discussed in detail in chapter II.

Such form-focused communication tasks represent an intrinsically motivating classroom technique and can be integrated alongside a more traditional approach to teaching grammar (R. Ellis, 2003). Teaching grammar through such activities is compatible with the philosophy of learner-centered language teaching and, at the same time, allows for the teacher's input and guidance as well as for corrective feedback or error treatment. This technique for teaching grammar is believed to promote the development of communicative fluency, which is the primary goal of language instruction, without sacrificing syntactic and morphological accuracy (R. Ellis, 2001; Larsen-Freeman, 2003).

Some target grammatical structures lend themselves especially well to being introduced within communicative settings set up by the teacher or the teaching materials when learners already know what they are trying to say but lack the means to do so

adequately in the TL (Long, 2007). Alternatively, these activities may take place after the targeted structure already has been introduced in some way and, occasionally, after it already has been practiced in more mechanical, teacher-controlled exercises, which is possible in a *task-supportive* (vs. strictly task-based) curriculum (R. Ellis, 2003). The rationale for having learners complete such communicative tasks through interaction with each other is based on the belief that this instructional technique offers the opportunity for more natural learning inside the classroom. It helps overcome the so-called *inert knowledge problem*, that is, the unfortunate situation when knowledge of the rule and ability to produce the correct form when prompted do not translate into ability to use it appropriately when the learner's primary attention is on conveying meaning (Larsen-Freeman, 2001b). Empirical research findings conclusively demonstrate that knowledge of grammar rules is not a guarantee that the learner will be able to use these rules for communication (Jackson, 2008; Nunan, 1999). Therefore, according to Lightbown (2007), any language feature that is taught didactically (i.e., outside of a natural communicative language-use setting) has to be practiced communicatively for the teaching to have any practical effect.

From the viewpoint of skill acquisition, tasks undoubtedly help learners progress from *declarative* knowledge about the target structure (i.e., the knowledge of the associated rule) to *proceduralized* knowledge (i.e., the skill of forming the structure) and, finally, to *automaticity* (i.e., fully automatized and implicit skills of using the structure appropriately; DeKeyser, 2007). The automaticity then allows for learner's attentional resources to be allocated to other aspects of the utterance (Skehan, 1998), for example, to meaning, discourse organization, pragmatics, lexis, and so on. Unlike traditional types of

grammar practice, task-based interaction constitutes *transfer-appropriate processing* of TL structures that DeKeyser defined as processing that is conducive to developing skills transferrable to real-use situations in TL speaking environments.

Although there appears to be a clear theoretical rationale for using task-based interaction in teaching grammar, there is a distinct disconnect in the minds of some language teachers and researchers between the ways in which TL grammar, on the one hand, and TL communication, on the other hand, are conceptualized and taught (Larsen-Freeman, 2001a). Teachers may not understand fully the principles of task-based instruction, their own role in designing appropriate form-focused tasks, and ways to facilitate learner interaction in these tasks effectively in class. Some researchers, most notably Seedhouse (1999, 2005), believed that learner-to-learner interaction only leads to fossilization of faulty grammatical structures in the learner's interlanguage (i.e., the developing implicit system; Long, 2006; Selinker, 1972). Sheen (1994, 2003) and Swan (2005) also considered learner interaction in small groups to be incompatible generally with effective teaching of grammar. Critics of the task-based approach to teaching TL features argue that there is no empirical evidence to demonstrate that it is superior to the traditional grammar teaching approaches largely relying on didactic nontask activities that analyze discrete grammar points outside of a communicative context (Long, 2000).

The idea of designing form-focused tasks for practicing grammar in the classroom faces the criticism from the other side of the language teaching beliefs spectrum as well. Krashen (1981, 1993), based on his claim that conscious language learning (vs. naturalistic language acquisition) never leads to interlanguage development, questioned whether any deliberate form-focused instruction can have more than a peripheral effect.

The beliefs of Krashen and his followers represent an extreme *noninterface* position that preempts any discussion of the effectiveness of deliberate focus on language form in the language classroom (Krashen, 1985).

There are distinct differences of opinion even among the most prominent supporters of focus on form within task-based language teaching. For example, Long's (2000) position mostly recognized brief diversion of the students' attention to problematic grammatical structures as an issue arises incidentally during completion of a real-world communicative task, that is, Long mostly supported learner-triggered, reactive (vs. proactive or preemptive) and incidental (vs. planned) attention to language form. In opposition to Long (2000), R. Ellis (2003) and Willis and Willis (2007), among many others, recognized the need for planned, deliberately designed grammar-focused interactive tasks. Moreover, R. Ellis, in particular, advocated the use of tasks alongside more traditional teaching and, more specifically, inclusion of more traditional activities and techniques in the pretask (i.e., planning and preparation for task completion) and the posttask (i.e., feedback and reflection) phases.

Although some empirical studies have provided evidence that task-based interaction can facilitate learner acquisition of specific TL features (Mackey, 1999), other studies did not find empirical support for the existence of such a relationship (Loschky, 1994). In view of such disparate research findings and stark differences of opinion over the role of planned grammar-focused tasks requiring learner-to-learner interaction, more systematic empirical evidence of their effectiveness is needed.

### Purpose of the Study

This meta-analytic study examined research into the effectiveness of classroom

task-based interaction that occurs during focused (structure-based) communication tasks as an instructional technique for improving mastery of specific TL forms. The purpose of this investigation was multifaceted: (a) to contribute to building a body of empirical evidence regarding the effectiveness of task-based interaction in acquisition of specific TL grammatical structures, (b) to investigate the impact of various moderator variables that influence the effectiveness of such interaction, for example, characteristics of the task used as the treatment, characteristics of the target structure, the degree of similarity between the learners' first language (L1) and L2, learner proficiency levels, and so forth, and (c) to define the best practices in task-based form-focused instruction in FL and L2 teaching based on the empirical evidence of differential effects of instruction-related moderator variables (if there is evidence of such differential effects). Additionally, the systematic examination of primary research studies in the domain allowed the meta-analyst to capture the current research trends and practices, point out areas needing improvement, and outline possible directions for future research.

The present meta-analytic study involved quasi-experimental and experimental studies where the treatment included teaching of FL and L2 grammar through interactive classroom practice activities that by design predispose learners toward using particular targeted structures repeatedly, but, unlike mechanical drills, require the learners to engage in exchange of real meaning. Such activities are referred to in this study as focused communication tasks. Even though the terms used to refer to this type of practice may vary in the SLA literature, all of these activities are similar in the following sense: (a) they combine focus on specific target structures with focus on meaning (Doughty, 2001; R. Ellis, 2001, 2003; Larsen-Freeman, 2001a, 2001b, 2003), (b) the learners are given a



nonlinguistic purpose for their interaction, for example, to solve a real-world problem; predict, negotiate, come up with a joint plan of action; and so forth, as opposed to drills where utterances are formed exclusively for language display purposes (R. Ellis, 2003; Leaver & Willis, 2004), and (c) there is an observable outcome, that is, the solution to a problem; prediction, plan, ranked list, schedule; and so forth (R. Ellis, 2003; Leaver & Willis, 2004).

Based on the overwhelming evidence that language acquisition processes in prepubescent children are entirely different from adult acquisition processes, at least in immersion-like environments (Curtiss, 1988; DeKeyser, 2000; Hyltenstam & Abrahamsson, 2006; Lightbown & Spada, 1993; Newport, 1990), the research domain was limited to primary studies that investigate acquisition of TL structures by adult learners (i.e., postpubescent learners who are 13 or more years old). The dependent variable(s) in the meta-analyzed primary studies was or were the students' acquisition (i.e., learning) of the target structure(s) as measured by the scores on immediate and, possibly, delayed posttests.

The effectiveness of task-based-interaction treatments used in the primary studies was assessed by means of the basic index for the effect-size value (Cohen's, 1977, *d*), that is, *standardized mean difference*. The effect-size values was calculated by subtracting the mean of the control or comparison group from the mean of the experimental (task-based-interaction) group and dividing the difference by the pooled standard deviation. For a subset of studies that investigated pretest to posttest score differences for a single group, Cohen's *d* was calculated as the *standardized mean gain* by dividing the mean gain value (i.e., the difference between the mean posttest and the

mean pretest scores) by the pooled standard deviation of the pre- and posttest values (Norris & Ortega, 2000). The resulting standardized-mean-gain effect-size values were not comparable with the standardized-mean-difference effect-size values and, therefore, were analyzed separately.

After the final sets of effect-size values were calculated and adjusted for bias (Hedges & Olkin, 1985), the individual effect sizes were averaged together (for the standardized-mean difference and for the standardized-mean gain) to depict the overall magnitude of the effects of task-based interaction on the students' acquisition of the target structure(s). Cooper (2003) warned against combining primary studies that use different types of participants and outcome measures within one meta-analysis and suggested that several separate meta-analyses be completed instead within the same research synthesis in order for the meta-analyst to be able to make summary statements about relationships between the variables. Following the established practice for research syntheses and meta-analyses in the field of language teaching and learning (Keck et al., 2006; Mackey & Goo, 2007; Norris & Ortega, 2000, 2006b), the overall mean effect size for the task-based interaction was interpreted as a suggestive (rather than definitive) finding. Differences between specific task-based-interaction treatments, participants, and outcome measures were treated as potential moderator variables that mediate effects of task-based interaction (i.e., multiple separate analyses were completed for subsets of studies that shared certain coded substantive or methodological characteristics). Because the aggregation, that is, the number of qualifying studies for various levels of the moderator variables, typically was small, the findings regarding the effects of moderator variables are presented primarily as descriptive.

Research findings suggest that the students' performance in oral communication tasks and the resulting learning of L2 features may be dependent on such variables as the type of task used as treatment as well as a whole range of other variables (R. Ellis, 2003; Gass & Mackey, 2007; Gass & Selinker, 2008; Keck et al., 2006; Long, 2007; Samuda, 2007; Van den Branden, 2006; Willis & Willis, 2007). These moderator variables may be related to specific characteristics of the learners (e.g., L1, proficiency level, age, etc.), instructional treatments (e.g., presence of explicit grammar instruction in the pretask or posttask stage), target TL grammatical structures (e.g., whether they are morphological or syntactic, simple or complex, etc.), and even study research designs (e.g., whether the participants have volunteered for the task-based-interaction group or not). To gain insight into these possible relationships, studies that shared each of these identified characteristics were meta-analyzed together, and the mean effect sizes were compared for different levels of these variables if there was sufficient aggregation of studies for each level. Various types of potential moderator variables are discussed in detail in chapter II. In those instances when the moderator variables were related to task design or teaching practices (e.g., the type of task, presence or absence of certain elements of instruction in pretask and posttask stages, etc.), after analyzing the impact of these variables, the meta-analyst attempted to present an overview of the best practices in using focused communication tasks to the extent possible.

There is considerable variation in the types of posttests used to measure TL acquisition (Keck et al., 2006; Mackey & Goo, 2007; Norris & Ortega, 2000, 2006b). Some research findings have suggested that the type of posttest used in the primary study to measure acquisition of the target grammatical structure may have an effect on the

students' scores (Erlam, 2003; Gass & Mackey, 2007; Norris & Ortega, 2000). It is possible that students who received traditional grammar explanations and drills perform better on a grammaticality-judgment or a fill-in-the-blanks test than on a test that involves oral production tasks. Conversely, students who received communicative grammar practice may be better prepared for assessment involving oral production than other types of tests. For this reason, the meta-analyst investigated what effect the type of outcome measure used to assess students' performance after a task-based-interaction treatment has on the findings of the study.

The research methodology that was used in the present meta-analytic study is discussed in greater detail in chapter III. There are certain challenges that face meta-analysts in the field of FL teaching and learning in addition to the issue of lack of uniformity of the teacher- and researcher-designed posttests that typically are used to measure acquisition of specific grammatical structures. Primary studies in the field frequently do not adhere to stringent criteria for research design and reporting (Lazaraton, 2000; Norris & Ortega, 2006a, 2006b). For this reason, some of the "classical" guidelines for a meta-analysis outlined by Cooper (2003) and Lipsey and Wilson (2001), among others, could not be followed in the present meta-analytic study. Additionally, as expounded by Norris and Ortega (2000, 2006b), strict adherence to some of the prescribed guidelines while investigating the effects of L2 instruction may result in obfuscation of important differences among the variables that precisely are the focus of the meta-analytic investigation. For example, primary study designs contrasting a single experimental condition with a single control condition that are ideal from the point of view of a meta-analysis are rare in FL teaching and learning, and multiple comparison

groups that receive a variety of instructional treatments typically are present (Norris & Ortega, 2006b). This consideration leads meta-analysts to a principled decision not to follow Lipsey and Wilson's (2001) recommendation to combine within-study effect sizes in order to avoid the problem of nonindependence of effect-size values when the goal is to investigate how specific characteristics of each treatment impact the effect of this treatment (Keck et al., 2006; Norris & Ortega, 2000). In those instances when an alternative strategy (i.e., not a "classical" prescribed strategy for meta-analyses) was followed based on Norris and Ortega's recommendations (2000, 2006b) for the SLA field, the rationale is provided in chapter III.

### Theoretical Rationale

Foreign language grammar has been viewed by some classroom teachers exclusively as a set of rigid prescriptive rules about what constitutes correct as opposed to incorrect structuring of utterances. Based on this conception of grammar, its teaching quite logically was understood to entail transmission of the knowledge of rules and intricacies of this system from the teacher to the student (Larsen-Freeman, 2003; Purpura, 2004).

Since the 1980s, the profession has been moving toward a more holistic view of grammar. The most comprehensive conceptualization of grammar has been provided by Larsen-Freeman (1995) who presented grammar as a higher order concept within linguistics with three interrelated dimensions: form, meaning, and use (i.e., situational appropriateness). According to Nunan (1999), this model attempted to integrate three aspects of linguistics that traditionally have been kept separate: syntax (i.e., study of form), semantics (i.e., the study of meaning), and pragmatics (i.e., the study of use).

Therefore, Nunan (1999) defined grammar as “the study of how syntax (form), semantics (meaning), and pragmatics (use) work together to enable individuals to communicate through language” (p. 101). This section briefly examines the theoretical frameworks for teaching grammar within CLT in light of the emphasis on learning the language for and through completing communicative functions. The two frameworks that are most important to the investigation of the role of task-based interaction in teaching TL grammar are task-based language teaching and Focus on Form.

### *Task-Based Language Teaching*

*Task-based language teaching (TBLT)* has been proposed as a method of promoting learning of form in the context of meaningful communication (R. Ellis, 2003; Long, 1997; Long & Crookes, 1993; Long & Robinson, 1998; Nunan, 1999; Skehan, 1998, 2001; Willis, 2004; Willis & Willis, 2007). A classroom language learning *task*, as opposed to an *exercise* or *free conversation*, is defined as an activity during which the learners’ attention is primarily on meaning, rather than form (Nunan, 1989); however, unlike free conversation in the TL, a task has a workplan (R. Ellis, 2003), presents a real-world communication problem to be solved, and is assessed in terms of its pragmatic outcome (Skehan, 1998). For example, learners can be asked to reach a consensus about a real-life issue, design a joint plan of action, predict the outcome of a situation, prepare a list of possible arguments against a proposition, report discrepancies between two sources of information, conduct a poll and report its results, and so forth. The concept of task and TBLT methodology are discussed in more detail in chapter II.

Empirical research findings have indicated that engaging in tasks can promote formal learning both when interaction takes place between native speakers (NS) and

nonnative speakers (NNS) of the TL (Mackey, 1999) as well as in NNS-NNS interaction (Adams, 2007; Williams, 1999). The biggest concern associated with TBLT is that learners will focus on meaning and let the language “drift by” (Lightbown, 2007). This concern leads to rejection of TBLT by some classroom practitioners, especially when teaching of grammar is involved. Cobb and Lovick (2007) reported that some language teachers hold a belief that TBLT can be useful in the development of TL fluency but not grammatical accuracy. In particular, the ability of communicative tasks to target acquisition of specific language structures is questioned. Although some empirical evidence of the effectiveness of TBLT in developing mastery of specific target structures has been reported (Keck et al., 2006), it remains scarce. The purpose of the present meta-analytic study was to expand the research domain in order to further the investigation of the effectiveness of learners’ task-based interaction. Long (1991, 2000) formulated the Focus on Form approach as a key methodological principle of TBLT that allows for teaching of TL grammar in the process of meaningful communication.

### *Focus on Form*

*Focus on Form (FoF)* is a feature of CLT that involves attention to linguistic features (e.g., morphological and syntactical) taking place in the context of performing a meaning-focused activity (Basturkmen, Loewen, & Ellis, 2004; Doughty & Williams, 1998; Long, 1991). It is differentiated from both *Focus on Forms (FoFS)* where grammatical features are extracted from context or communicative activity and are practiced in isolation in drill-like exercises and *Focus on Meaning (FoM)* where learners merely engage in communication using the language means they already have and no

attention to language form ever is intended deliberately (Long, 1991, 1996; Long & Robinson, 1998).

The term *form* sometimes is used to refer to all formal aspects of the language to include phonology (i.e., correct pronunciation), lexis (i.e., accuracy in using vocabulary items), pragmatics (i.e., situational appropriateness and accuracy in conveying the intent of one's message), discourse-organization features, and so forth (Doughty & Williams, 1998). In this study, *form* is used only to refer to grammatical aspects of the language such as morphology (i.e., word form changes used to mark grammatical categories of number, gender, person, case, tense, voice, aspect, transitivity, etc.) and syntax (i.e., patterns for combining sentences, sentence clauses, and parts of clauses).

Just as Norris and Ortega's (2000) meta-analysis of effectiveness of L2 instruction, this study does not use the original, restrictive definition for FoF as brief diversion of learners' attention to form only as a reactive, learner-triggered activity (Long & Robinson, 1998) but includes planned, proactive attention to form as long as it meets the criteria for integration of teaching of form and real-world communicative tasks. The definition of FoF adopted in this study is not as broad as R. Ellis' (2001) definition of *Form-Focused Instruction (FFI)* that refers to any planned or incidental activity whose purpose is to induce learners to pay attention to form regardless of its nature (i.e., regardless of whether it is communicative or traditional in nature).

In the classroom, FoF can be accomplished in a variety of ways, both through implicit and explicit means. Implicit means include recasts, that is, more correct reformulations of the learner's utterance (Lyster & Ranta, 1997), as well as clarification requests, comprehension checks, confirmation checks, repetitions, and so forth. These



techniques seek to direct the learners' peripheral attention to form without diverting their focal attention away from meaning. As compared with implicit techniques, explicit techniques such as explicit error correction and metalinguistic feedback engage the learners' focal attention (Doughty, 2001). Task-based interaction that occurs in focused communication tasks can be considered an FoF technique because such activities have a nonlinguistic real-world goal for the learners' interaction with other task participants, yet at the same time they are designed to improve control over specific grammatical forms, provided that appropriate monitoring and feedback take place. The effectiveness of task-based interaction as an FoF instructional technique has been investigated in a limited manner as explained in the subsequent section titled *Background and Need*. The purpose of the present study is to expand this investigation. The FoF approach in FL and L2 teaching is reviewed in more detail in chapter II.

### Background and Need

Meta-analysis is still a relatively new research methodology in the field of applied linguistics and SLA (Norris & Ortega, 2006b). Nevertheless, a number of meta-analytic studies have been completed in the 2000s that investigated effectiveness of TL instruction in general or, more specifically, effectiveness of interaction in TL acquisition (Jeon & Kaya, 2006; Keck et al., 2006; Norris & Ortega, 2000; Mackey & Goo, 2007; Plonsky, 2010; Russell & Spada, 2006; Spada & Tomita, 2010). Three previous meta-analyses that are related most closely to the topic of the present study are reviewed briefly in this section. Their limitations in view of the purpose of the present study and the need for further research are provided in a separate subsection titled *Limitations of the Three Previous Meta-Analyses*.

*Norris and Ortega's Research Synthesis and Meta-Analysis  
of Effectiveness of L2 Instruction*

Norris and Ortega (2000) employed systematic procedures for research synthesis and meta-analysis to summarize findings from experimental and quasi-experimental investigations into the effectiveness of different types of L2 instruction published between 1980 and 1998. (The meta-analysts did not focus specifically on the effectiveness of TBLT.) Comparisons of the average effect sizes from 49 unique studies indicated that focused L2 instruction (i.e., instruction in specific targeted grammatical and lexical language items) leads to large gains. The mean effect size for L2 instruction across all instructional treatments was  $d = .96$  ( $SD = .87$ ) on immediate posttests.

The meta-analysts compared mean effect sizes for explicit versus implicit instructional techniques and demonstrated that explicit techniques on average resulted in greater gains ( $d = 1.13$ ,  $SD = .86$ ) than implicit techniques ( $d = .54$ ,  $SD = .74$ ; Norris & Ortega, 2000). Similarly, Spada and Tomita (2010) who conducted a meta-analysis investigating the effectiveness of explicit over implicit instruction for simple and complex structures reported that the results indicated larger effect sizes for explicit over implicit instruction for both types of structures. Norris and Ortega also found that techniques that fall under the FoF approach (i.e., techniques that briefly focus on language features within meaningful communicative activities conducted in the TL) were equally as effective as those that fall under the FoFS approach (i.e., techniques that focus on language features outside of a communicative context). The mean effect sizes were  $d = 1.00$  ( $SD = .75$ ) for FoF and  $d = .93$  ( $SD = .96$ ) for FoFS. Because FoF techniques provide the students with opportunities for practice in processing input for meaning, communicating their own meaning, or both, this finding may be interpreted tentatively as

an indication of greater benefits of FoF techniques. Task-based interaction that occurs in focused communication tasks that was investigated in the present study is an explicit FoF technique.

Norris and Ortega (2000) reported that primary researchers employed a variety of different outcome measures (i.e., posttests measuring acquisition of target L2 features) as dependent variables to evaluate the effectiveness of instructional treatments. These outcome measures ranged from discrete-point tests that prompted examinees to display grammatical knowledge to free-oral-production tasks where the examinees' performance on these tasks was coded and analyzed in different ways. Norris and Ortega calculated mean effect sizes for four main posttest types: (a) constrained-constructed-response measures used in 65% of the studies ( $d = 1.20$ ,  $SD = .95$ ), (b) selected-response measures used in 39% of the studies ( $d = 1.46$ ,  $SD = 1.23$ ), (c) metalinguistic-judgment measures used in 29% of the studies ( $d = .82$ ,  $SD = .79$ ), and (d) free-constructed response used in 16% of the studies ( $d = .55$ ,  $SD = .97$ ). These findings showed that the mean effect sizes associated with metalinguistic-judgment tests and free-constructed responses were substantially lower than for selected responses or constrained-constructed responses. As the meta-analysts pointed out, study findings varied by as much as .91 standard deviation units depending on the type of outcome measure(s) used. Because the 95% confidence intervals for all four types of outcome measures overlapped, no inferences could be made. Such substantial variability in study outcomes based on the type of posttest used and even within one posttest type, as reported by Norris and Ortega, warrants further investigation. Different types of posttests that traditionally are used in primary research to

measure acquisition of grammatical TL items and associated measurement issues are discussed in detail in chapter II.

Due to a rather broad nature of their research purpose, Norris and Ortega (2000) reviewed a wide range of primary studies. Among those reviewed were studies that investigated whether learners' metalinguistic awareness of specific L2 forms facilitated acquisition (Fotos & Ellis, 1991; Swain, 1998), whether negative feedback was beneficial for L2 development, and, if so, what types of feedback were more effective (Carroll & Swain, 1993; White, 1991), whether comprehension practice was as effective for learning L2 features as production practice (DeKeyser & Sokalski, 1996; Salaberry, 1997), and so forth. In other words, Norris and Ortega's meta-analysis did not focus specifically on face-to-face task-based classroom interaction. Furthermore, some of the reviewed studies involved computer-mediated instruction rather than face-to-face instruction (DeKeyser, 1997; Nagata, 1993, 1998).

Consequently, the primary studies meta-analyzed by Norris and Ortega (2000) involved a wide variety of instructional techniques, for example, input flooding (i.e., providing learners with texts where the target structure abounds), textual enhancement (i.e., typographical enhancement of the target structure in the input such as color-coding, bolding, italicizing), recasts (i.e., native-like reformulations of the learners' utterances), consciousness-raising activities, input practice, output practice, metalinguistic practice, and so forth, with a total of about 20 subtypes of instructional techniques. The subtype labeled *output practice* represented traditional exercise-like practice. Task-based interaction, which is the focus of the present study, was not identified as a specific instructional technique. Based on the classification created for the purpose of the meta-

analysis, Norris and Ortega (2000) coded the independent variable in 2 of the 49 included studies involving task-based interaction as *interactionally modified input* (Loschky, 1994; Mackey & Philp, 1998). A preliminary examination of the primary studies included in Norris and Ortega's meta-analysis has revealed that, in addition to these two studies, possibly only one additional study (Long, Inagaki, & Ortega, 1998) out of all 49 meta-analyzed studies involves an instructional treatment that may qualify to be considered task-based interaction. Therefore, an investigation that included a greater number of studies that specifically focus on the effectiveness of task-based interaction in form-focused tasks was required.

*Keck, Iberri-Shea, Tracy-Ventura, and Wa-Mbaleka's Meta-Analysis Investigating the Empirical Link Between Task-Based Interaction and Acquisition*

Keck et al. (2006) conducted a meta-analysis of 14 studies published between 1980 and 2003 investigating the effectiveness of form-focused communicative tasks as an instructional technique for improving both structural and lexical accuracy. Therefore, in contrast to Norris and Ortega's (2000) work, their meta-analysis narrowly focused on one specific FoF technique for the development of specific L2 forms, that is, meaningful task-based interaction, which is also the focus of the present study.

The meta-analysts reported large main effects of task-based interaction on acquisition for L2 target items  $d = .92$  ( $SD = .68$ ) on immediate posttests. Specifically, the mean effect size  $d = .94$  ( $SD = .67$ ) was reported for grammatical items and  $d = .90$  ( $SD = .75$ ) for lexical items (Keck et al., 2006). These results represented large effects based on Cohen's (1977) suggested guidelines for interpretation of the magnitude of effect sizes. The mean effect sizes for grammatical and lexical items together were even larger for short-delayed posttests ( $d = 1.12$ ) and long-delayed posttests ( $d = 1.18$ ) than for

immediate posttests. The magnitude of change from pretests to posttests for the 5 out of 14 studies that reported both the pretest and posttest scores was also large for the treatment groups  $d = 1.17$  ( $SD = .87$ ) as compared with the medium mean effect size of  $d = .66$  ( $SD = .55$ ) for the control and comparison groups, even though the 95% confidence intervals overlapped.

Keck et al. (2006) attempted to investigate the possible effect of the task type on the dependent variable, for example, score on a measure of acquisition of the target L2 feature. Different task types based on various classifications are provided in chapter II. Keck et al. were able to calculate the mean effect sizes only for *jigsaw tasks* (i.e., tasks in which the input material is divided between the participants so that they all are required to exchange information with each other in order to complete the task successfully; R. Ellis, 2003; Pica, Kanagy, & Falodun, 1993), *information-gap tasks* (i.e., tasks in which one participant holds information that the other participants do not have and that needs to be provided to them in order for the task to be completed successfully; R. Ellis, 2003; Prabhu, 1987), and *narrative tasks* (i.e., tasks that require participants to produce a narration, such as about a past event; Keck et al., 2006). The mean effect-size values were  $d = .78$  for jigsaw tasks,  $d = .91$  for information-gap tasks, and  $d = 1.60$  for narrative tasks; however, the calculation of the mean effect size for the narrative tasks was only based on two treatments. Additionally, the 95% confidence intervals overlapped for all the three types. Mean effect sizes for other types of tasks could not be reported (e.g., opinion-gap tasks, problem-solving tasks, etc.). Keck et al. did not attempt to investigate the effects of task types based on other known classifications provided in chapter II of this study (e.g., so-called closed vs. open tasks based on the number of potential

acceptable solutions). The meta-analysts did not report whether they had investigated the effect of the type of outcome measure used as the posttest on the findings of primary studies.

Grammatical structures that are the focus of the present study were the targets of instruction only in 7 out of 14 instructional treatments in the primary studies meta-analyzed by Keck et al. (2006). Additionally, as is frequently the case in laboratory studies, all but 3 of the 14 meta-analyzed studies involved learners interacting with the researchers, teachers, and other NSs such as teaching assistants (TAs) or tutors, rather than with NNSs, while completing the communicative tasks. As explained in the subsection titled *Limitations of the Three Previous Meta-Analyses*, a greater aggregation of studies that involve learner-to-learner interaction was desirable. Because the research purpose of Keck et al.'s meta-analysis is related closely to the research purpose of the present meta-analytic study, a detailed review of Keck et al.'s report, including search procedures, data analysis procedures, and findings, is provided in chapter II.

#### *Mackey and Goo's Research Review and Meta-Analysis of Interaction Research*

Mackey and Goo (2007) investigated empirical research into effects of interaction on acquisition of both grammatical and lexical TL features published between 1990 and June 2006. Mackey and Goo considered their meta-analysis to be an update to the work conducted by Keck et al. (2006), who meta-analyzed interaction research up to the year 2003, as well as to the meta-analysis conducted by Russell and Spada (2006), who focused on the contribution of corrective feedback to L2 acquisition through the same year. Because Russell and Spada identified their research purpose as the investigation of effectiveness of corrective feedback, that is, error correction, rather than the effectiveness

of interaction, their meta-analysis was not reviewed.

Having meta-analyzed 28 studies, Mackey and Goo (2007) concluded that interaction plays a strong facilitative role in the learning of both lexical and grammatical items in short term and long term compared with little or no interaction. Mackey and Goo reported the following mean weighted effect sizes for treatment groups across eligible studies: on immediate posttests  $d = .59$  ( $SD = .61$ ) for grammar,  $d = 1.32$  ( $SD = .75$ ) for lexis; on short-delayed posttests (i.e., 7 to 29 days after the treatment)  $d = 1.07$  ( $SD = .82$ ) for grammar,  $d = .85$  ( $SD = .59$ ) for lexis; and on long-delayed posttests (i.e., 30 or more days after the treatment)  $d = .99$  ( $SD = .69$ ) for grammar and  $d = .96$  ( $SD = .04$ ) for lexis. The meta-analysts pointed out that, as can be seen from these results, the effect of interaction for grammar on immediate posttests ( $d = .59$ ) could be interpreted only as medium based on Cohen's (1977) suggested guidelines for interpretation of the magnitude of effect sizes. The 95% confidence intervals for this mean effect size for grammar and for the mean effect size for lexis on immediate posttests did not overlap, which indicated a statistically significant difference (alpha level = .05). This finding was not consistent with Keck et al.'s (2006) findings of no difference in effect sizes between lexis and grammar. For the subset of studies that reported within-group changes, that is, learners' gains between the pretests and the posttests, Mackey and Goo found a large mean effect size for the interaction groups  $d = 1.09$  ( $SD = .93$ ) compared with the mean effect size for control and comparison groups  $d = .44$  ( $SD = .42$ ). Unlike in Keck et al.'s (2006) meta-analysis, the 95% confidence intervals did not overlap for these two mean effect sizes.



Mackey and Goo (2007) used a different classification system for the types of dependent measures used in primary studies to measure acquisition of target FL and L2 items than the classification that Norris and Ortega (2000) used in their meta-analysis. Mackey and Goo reported that, on immediate posttests, so-called prompted-response measures that represented a combination of metalinguistic-judgment responses and selected responses in Norris and Ortega's classification were associated with small gains ( $d = .24$ ,  $SD = .56$ ). Open-ended prompted-production measures that were equivalent to Norris and Ortega's free-constructed response were associated with medium gains ( $d = .68$ ,  $SD = .52$ ), whereas closed-ended prompted-production measures equivalent to Norris and Ortega's constrained-constructed response were associated with large gains in acquisition ( $d = 1.08$ ,  $SD = .93$ ). These differences were statistically significant, that is, the 95% confidence intervals did not overlap. These findings are different from Norris and Ortega's findings regarding the effects of the type of outcome measure on the findings of the primary studies. Mackey and Goo's findings regarding the effects of test type for short-delayed and long-delayed tests were inconclusive. An examination of the effect of test types on the outcomes of primary studies investigating the effectiveness of focused communication tasks targeting acquisition of grammatical structures also was conducted in the present meta-analysis.

Similar to Russell and Spada's (2006) meta-analysis of effectiveness of corrective feedback, Mackey and Goo (2007) focused their attention on the effectiveness of corrective feedback that occurs during interaction. The meta-analysts included studies that utilized communication tasks either as the treatment or as a way of providing context for other treatments in question to occur. These meta-analysts explicitly excluded studies

that did not deal with one of the three types of corrective feedback (i.e., error correction) such as (a) corrective recasts (i.e., more target-like reformulations of the learner's utterances), (b) negotiation including comprehension checks and clarification requests, and (c) metalinguistic feedback. Mackey and Goo included several studies investigating corrective feedback that were not included in Keck et al. (2006) because Keck et al. focused primarily on the effectiveness of task-based interaction as instructional treatment.

Unlike Keck et al. (2006), Mackey and Goo (2007) included studies that investigated child L2 acquisition and studies that employed synchronous computer-mediated interaction. The focus of the present study is on adult language acquisition in face-to-face contexts. On a final note, Mackey and Goo excluded studies (Adams, 2007) that used so-called custom-made posttests that were custom-designed by researchers for each learner based on the errors the learner made in the use of target structures during interaction with a purpose of measuring this particular learner's individual learning. Upon examination of Adams' study, the meta-analyst concluded that it met the inclusion criteria outlined in chapter III and therefore included it in the present meta-analysis. It was anticipated that the inclusion of studies employing custom-made tests may result in a larger accumulation of primary studies and thus help investigate effectiveness of task-based interaction in more depth. Additionally, it was hoped that if more eligible primary studies utilizing custom-made posttests were located, it would allow the meta-analyst to investigate the effect of the type of posttest used as the outcome measure on the findings of the study.

#### *Limitations of the Three Previous Meta-Analyses*

As pointed out earlier, because Norris and Ortega's (2000) meta-analysis was

focused broadly on effectiveness of explicit versus implicit instructional techniques as well as FoF versus FoFS techniques, these researchers did not investigate the effectiveness of task-based interaction specifically. Moreover, the treatments whose effects they sought to compare ranged from one hour in duration to multiple weeks in duration and could include one instructional technique or a combination of multiple techniques. In the latter case, some of such compound treatments may have included task-based interaction, but it would be impossible to isolate its effects.

Mackey and Goo's (2007) meta-analysis focused specifically on the effectiveness of corrective feedback that occurs during such interaction and excluded studies that did not focus explicitly on corrective feedback. Investigation of effectiveness of different types of corrective feedback, or error correction, is a large, widely-researched subfield within the field of SLA (Ellis, Loewen, & Erlam, 2006; Lyster & Ranta, 1997; Russell & Spada, 2006), but it was not the focus of the present meta-analytic study. Moreover, corrective feedback occurs to a larger degree if the learners interact with a teacher, a teaching assistant, or other NS interlocutor. Therefore, in a study, it typically happens under so-called laboratory conditions when the researcher can provide NS volunteers or TAs as partners to each learner or to each small group of learners. Arguably, dyadic or small-group learner-to-learner interaction among NNS under actual classroom conditions involves negotiation of meaning as well, but the error correction aspect of it may be very different in purpose, quantity, and quality from the error correction provided by a teacher-type interlocutor. From the pedagogical perspective, it is important to employ small-group classroom tasks that are beneficial for the development of mastery of L2 features in the absence of NS assigned to each group. Therefore, the present meta-

analytic study did not use the presence of focus on corrective feedback as an inclusion criterion.

Additionally, Mackey and Goo (2007) included child studies even though language acquisition processes in adults have been shown to be different from language acquisition processes in children due to so-called maturational constraints (DeKeyser, 2000; Harley, 1986; Hyltenstam & Abrahamsson, 2006). Both Norris and Ortega (2000) and Mackey and Goo included investigations of written computer-mediated communication (CMC) that is independent of the immediate face-to-face classroom context (Wildner-Bassett, 2005) and, therefore, has characteristics that are quite different from oral face-to-face interaction (Cuskelly & Gregor, 1994; O'Rourke, 2005). As Keck et al. (2006) stated, both lowering the learner age requirement below postpubertal adolescents and including CMC studies is likely to introduce additional confounding variables that are not desirable.

Only Keck et al. (2006) specifically investigated the effects of oral task-based interaction on mastery of specific TL items by adult learners; however, the target linguistic features for the treatments in the primary studies included both grammatical and lexical features. The present study is focused exclusively on learners' acquisition of grammatical structures because acquisition processes for grammar are believed to be quite different from those involved in the acquisition of lexis. In this respect, Mackey and Goo (2007) themselves pointed out that "some interaction researchers have suggested that there may be a delayed effect of interaction on grammar which takes longer to become effective but it is then durable" (p. 439).

The majority of the studies included in all three meta-analyses involved NS-NNS interaction. As stated earlier, all but three studies meta-analyzed by Keck et al. (2006) fell under this category. NS-NNS interaction (i.e., teacher-learner, NS tutor-learner, or researcher-learner interaction) investigated in most meta-analyzed studies is more likely to happen under experimental laboratory conditions than in a realistic classroom situation with many learners and one teacher present. Van den Branden (2007), among others, advocated strongly for classroom-based research in naturalistic FL and L2 language-learning settings. Therefore, in order to increase the fidelity of the treatment condition, it was essential to attempt locate and include in the present meta-analysis more studies investigating the effectiveness of learner-to-learner task-based interaction that targets acquisition of specific grammatical language items. This consideration was one of the main reasons for the decision to include unpublished studies in this meta-analysis as explained in chapter III in the *Data Sources and Search Strategies* section.

More studies of the effects of task-based interaction have been published since June 2006 (i.e., after the end of the timeframe for Mackey and Goo's [2007] meta-analysis) such as Toth (2008). It also was considered feasible that more relevant studies were published between 2003 (i.e., the end of the timeframe for Keck et al.'s [2006] meta-analysis) and 2006 that involved task-based interaction but did not focus on corrective feedback that was the focus of Mackey and Goo's investigation. The reasoning behind the present meta-analysis was that, if such studies are located and included in a new meta-analysis, it may be possible to conduct a more meaningful examination of certain moderator variables (e.g., type of task used as treatment, type of target structure,

presence of explicit instruction in the pretask and posttask phase, etc.) when these studies are meta-analyzed together with the previous ones.

Additionally, a publication bias resulted because none of the three meta-analyses reviewed here included studies from unpublished sources. The researchers themselves warned the readers of the potential for serious publication bias influencing the results of their meta-analyses. Keck et al. (2006) provided the following two reasons for their decision to include only published studies:

1. The meta-analysts wanted the readers to be able to compare the results of their meta-analysis with the more traditional narrative or vote-counting reviews, and these typically only take into account published studies.

2. Because unpublished studies are difficult to locate, Keck et al. (2006) believed that it almost is impossible to retrieve all relevant unpublished studies. The meta-analysts explained that they were concerned about the possibility of collecting an idiosyncratic and biased sample of fugitive literature.

This concern certainly was not without justification, and Mackey and Goo (2007) even mentioned two widely cited unpublished studies that had proved to be impossible to obtain. Nevertheless, such an a priori exclusion of unpublished studies makes the findings of the meta-analyses generalizable only to the top tier of the published professional literature. It was hoped that inclusion of unpublished studies, most importantly, doctoral dissertations and conference reports, in the present meta-analysis may open up more opportunities for the examination of relevant moderator variables.

Mackey and Goo (2007) excluded studies that used custom-made posttests designed based on the errors the learner originally made in order to be able to measure

this particular learner's individual learning (Kowal & Swain 1994; Swain & Lapkin 1998, 2001, 2002). Understandably, these meta-analysts were concerned that the use of such outcome measures may make comparisons "unrealistic," akin to comparing "apples and oranges." A counterargument can be made, however, that there is already a large amount of variation present between different types of posttests used in the candidate studies, so a priori exclusion of custom-made tests that represent quite an elegant technique of measuring learner-specific learning is not necessary under the circumstances. Norris and Ortega (2000) who, in their seminal report, meta-analyzed a large number of studies involving a very wide range of explicit and implicit techniques for teaching L2 form and a wide range of outcome measures certainly used comparisons that would be considered "unrealistic" under this point of view. Their approach is defensible, however, due to the fact that their purpose was to compare the effectiveness of explicit techniques in general with the effectiveness of implicit techniques.

On a final note, because there is no single definition of a classroom task and no complete agreement about what task-based language teaching (TBLT) entails in the field of FL and L2 teaching, it was important to include a working definition of a classroom task in this new meta-analysis, including delineating criteria that distinguish a task from a nontask classroom activity (see the *Definition of Task* section in chapter II). Keck et al. (2006) defined task-based interaction as conversational interaction in the TL that takes place among NNS learners of this language or between NNS learners and NS interlocutors (in pairs or small groups) while completing assigned oral communication tasks. The researchers used Pica et al.'s (1993) definition of *tasks* as activities that engage a pair (or a small group) of learners in work toward a particular goal. In absence of

elaboration, this definition may be interpreted and applied rather broadly (i.e., in reference to a large number of classroom activities that are not necessarily tasks as defined in the present meta-analysis). Considering the fact that there is no complete agreement in the fields of FL and L2 teaching and research about the meaning of the term *task*, it may be argued that the meta-analysts did not provide a sufficiently detailed explanation of how they operationalized tasks.

For example, the instructional treatment in Long, Inagaki, and Ortega's (1998) study that was included in all three meta-analyses reviewed here does not constitute an oral-communication task as defined in the present meta-analysis because it involved learners attempting to name objects in Japanese (e.g., "large red paper") using appropriate grammar and then hearing the NS interlocutor do it correctly (i.e., corrective recast). It is questionable whether this minimally contextualized activity in which learners simply named objects and did not produce any utterances that convey novel personal meaning would meet the criterial features for task as defined, for example, by R. Ellis (2003). The mere fact that NS interlocutors had to identify the appropriate piece of paper in their own set (i.e., task product, or outcome) may not be sufficient to qualify this activity to be deemed an oral-communication task.

Similarly, the small-group interaction in Garcia and Asencion's (2001) study included in both Keck et al. (2006) and Mackey and Goo (2007) simply involved individuals reviewing their notes together before they reconstructed the text that they had heard and then answering comprehension questions based on the text individually. The transcripts of the interaction provided in the study report showed that participants repeated what they had heard using not only the TL but also frequently their first



language (L1) to verify comprehension with one another. Both the object naming activity in Long et al.'s (1998) study and the text reconstruction activity in Garcia and Asencion's (2001) study, although undoubtedly representing a step forward from fill-in-the-blanks type exercises targeting the same grammatical forms, hardly meet the criteria for real-world communication tasks that typically require deeper levels of processing of information than labeling or recalling that are considered to be low order cognitive processes (Anderson & Krathwohl, 2001; Bloom, 1956). Keck et al. reported that they had not planned to evaluate the effectiveness of interaction that occurs within focused communication tasks specifically. Their goal was to investigate the effectiveness of interaction in learning specific TL features; however, they concluded at the end of their investigation that all such interaction described in the included primary studies occurred in tasks. The intent of the present study was to focus on structure-based communication tasks that meet the rigorous criteria defined in the SLA field, most notably by R. Ellis (2003) as interpreted by the meta-analyst. The requisite criterial features of tasks are reviewed in chapter II (see *Criterial Features of Tasks*).

### Research Questions

The following are the research questions that the present meta-analysis addressed:

1. To what extent is oral task-based interaction that occurs in focused (structure-based) communication tasks (in FL and L2 instruction of adult learners) effective (i.e., how large is the standardized-mean-difference effect size resulting from task-based interaction treatments compared with other types of grammar instruction for the learners' acquisition of the target grammatical structure)?
2. Is the standardized-mean-gain effect size (i.e., effect size based on the pre- to

posttest differences) larger for task-based interaction treatments as compared with other types of grammar instruction?

3. Is there a difference in effect-size values based on the type of focused communication task (e.g., information-gap vs. opinion-gap, closed vs. open, etc.) used in the task-based interaction treatment?

4. Is there a difference in effect-size values based on other factors such as the type of grammatical structure targeted by the task-based-interaction treatment, duration of instruction as well as miscellaneous other teacher-related, learner-related, and contextual variables?

5. Is there a difference in effect-size values based on what type of outcome measure (i.e., posttest measuring acquisition of the target grammatical structure) was used in the primary research study (e.g., metalinguistic judgment vs. selected response vs. oral-communication task)?

### Significance of the Study

This meta-analytic study has implications for FL and L2 teachers, curriculum developers, teacher educators, and language program supervisors. The pedagogical implications are related to the effectiveness of TBLT and possibilities of using tasks in improving the teaching of TL grammatical structures. Integrating the teaching of formal features of the TL with the development of communicative skills is a state-of-the-art instructional technique that is misunderstood or not accepted by some language teachers (Cobb & Lovick, 2007). It was, therefore, important to synthesize up-to-date empirical data that provide evidence of its effectiveness. This meta-analysis systematically evaluated the findings of four eligible previously analyzed studies together with the

findings of the studies that were not included in previous meta-analyses for reasons presented in this chapter as well as of the studies that appeared after June 2006, that is, after the end parameter for Mackey and Goo's (2007) search.

Because the search procedure included more sources of published reports as well as the so-called fugitive literature, the present meta-analysis includes three new primary studies that involved learner-led (i.e., NNS-NNS) interaction. (A fourth study involved some of the participants interacting with their NS teacher and other participants interacting with each other; however, the reported results did not differentiate between the two conditions.) Given the scarcity of data for the effects of learner-led task-based interaction, gaining more insight into the issue is crucial for understanding what happens specifically in learner-to-learner interaction in classroom settings as opposed to laboratory settings in which students complete tasks through interaction with teachers, TAs, or other NS interlocutors.

Previous meta-analyses that synthesized primary research study findings have provided some evidence of effectiveness of task-based interaction in learners' morphosyntactic development (Keck, 2006; Mackey & Goo, 2007). The present meta-analysis adopted a somewhat different perspective from one or both of the related meta-analyses through the following features: (a) exclusion of studies that focus only on effects of corrective feedback, (b) inclusion of both published and unpublished studies to expand the research domain, (c) imposing of more stringent criteria for oral-communication tasks, (d) focusing on adult (vs. child) learners and face-to-face (vs. computer-mediated) interaction, and so forth.

The results of the 15 published and unpublished studies included in the present meta-analysis were compared with Keck et al.'s (2006) and Mackey and Goo's (2007) findings and also interpreted in light of other meta-analytic findings where applicable, for example, in light of Norris and Ortega's (2000) findings regarding the effects of FoF versus FoFS instruction, Spada and Tomita's (2010) findings regarding the interactions of the target structure complexity and the explicitness of instruction, and Plonsky's (2010) findings regarding the effects of study quality on outcomes.

The present meta-analysis provided additional data regarding the effectiveness of task-based interaction as compared with no focused instruction in the target structure and as compared with other types of grammar-focused instruction (including traditional grammar instruction, input processing activities, etc.). Data related to within-group gains were presented and analyzed as well. In addition to overall weighted mean effect sizes, effect sizes for various moderator variables, and determinations of statistical significance (through the analysis of 95% confidence intervals), the analog to the (one-way) analysis of variance (ANOVA) was used to learn whether the moderator variables could account for the variability in the effect sizes. The latter part of the analysis (i.e., analog to the ANOVA) was not performed in previous related meta-analytic studies. Additional evidence in support of Mackey and Goo's hypothesis regarding the durability of effects of task-based interaction for grammatical structures was provided.

The addition of new eligible study reports, including reports of studies completed as dissertations that typically present more details, expanded the scope for the meta-analysis, allowed for a more comprehensive research synthesis, and resulted in a somewhat greater accumulation of studies sharing some of the already examined as well

as new moderator variables. For example, additional task characteristics such as open-endedness and convergence (see sections *Closed and Open Tasks* and *Divergent and Convergent Tasks* in chapter II) as well as characteristics of the target structures (see section *Types of the Target Structure* in chapter II) not addressed in Keck et al.'s (2006) and Mackey and Goo's (2007) meta-analyses, were analyzed in the present study.

In addition to outlining relevant pedagogical implications, this meta-analytic study contributes to capturing research and reporting practices in investigating effectiveness of specific instructional techniques in developing mastery of target L2 features and provides recommendations for continuous improvement of these practices. In particular, it delineates which research recommendations presented in Norris and Ortega's (2000) seminal meta-analysis have led to improvements in primary research practices in the field and points out areas where improvements may still be needed.

### Definition of Terms

As stated in previous sections of this chapter, there is no unanimity in the SLA field and the field of FL and L2 teaching regarding definitions of some key terms. The following are the definitions of key terms used in the present study. A list of additional definitions that may assist the reader in understanding some relevant issues discussed in the study is provided in the Appendixes (see Appendix B).

Acquisition is the process by which humans learn a second or foreign language in addition to their native language (Doughty & Long, 2006).

Acquisition of a target language item (i.e., a grammatical structure) is the degree of mastery of this language item demonstrated by a foreign or second language learner that is generally determined by the rate and accuracy of the use of this language item

(Doughty & Long, 2006). In this study, acquisition of a grammatical structure is the dependent variable operationalized as the score on a posttest designed to measure the degree of mastery of this grammatical structure.

Adult (foreign language or second language) learner is operationalized in this study as a learner who is 13 years of age or older. This operationalization is based on a widely supported claim that language acquisition processes in children who have not reached puberty (i.e., prepubescent learners) are different than in older learners (i.e., postpubescent learners). The so-called maturational (i.e., age-related) constraints for language learning to native-like levels are believed to manifest themselves approximately at the onset of puberty (i.e., around the age of 13; Hyltenstaam & Abrahamsson, 2003).

Analog to analysis of variance (ANOVA) is a statistical procedure used in meta-analyses to test the ability of a moderator categorical variable to explain the excess variability of the effect size discovered by means of a homogeneity test (i.e.,  $Q$  statistic; see Test of homogeneity). The  $Q$  statistic is subdivided into  $Q_{BETWEEN}$ , or  $Q_B$ , that represents the variance in effect sizes accounted for by the moderator variable, and  $Q_{WITHIN}$ , or  $Q_W$ , that represents within-group error. When the  $Q_B$  is statistically significant and the  $Q_W$  is not statistically significant, the moderator variable successfully accounts for the variability in effect sizes (Lipsey & Wilson, 2001).

Between-group contrast is a contrast between the performance of the experimental group and the control or comparison group on a posttest. It can be expressed by the standardized-mean-difference effect size (Lipsey & Wilson, 2001; see Standardized-mean-difference effect size).

Criterion features of a task are requisite characteristics that qualify a learning activity to

be considered a “task” as this term currently is defined in the field of second and foreign language teaching (R. Ellis, 2003).

Effect size is a common metric used to compare and interpret results from different studies. In a meta-analysis, effect sizes are extracted from numerical or statistical data provided in each included primary study. The basic index for the effect-size (Cohen’s, 1977,  $d$ ) was calculated by subtracting the mean of the control or comparison group from the mean of the experimental (task-based interaction) group and dividing the difference by the pooled standard deviation (see Standardized-mean-difference effect size). For a subset of studies that investigate pretest to posttest score differences for a single group, Cohen’s  $d$  was calculated by dividing the mean gain value (i.e., the difference between the mean posttest and the mean pretest scores) by the pooled standard deviation of the pre- and posttest values (Norris & Ortega, 2000; see Standardized-mean-gain effect size). The resulting standardized-mean-gain effect size is not comparable with the standardized-mean-difference effect size and, therefore, was treated separately. Cohen’s  $d$  values were converted to Hedges’s  $g$  values, which are unbiased estimates (Hedges & Olkin, 1985, p. 81).

Exercise is a learning activity that involves manipulation of language form and does not meet the characteristics of a task as it currently is defined in the field of foreign and second language teaching (R. Ellis, 2003).

Focus on Form (FoF) is an approach to instruction that draws learners’ attention to linguistic form, or features of the language, while the primary focus of their attention is on meaning (Doughty & Williams, 1998; Long, 2000).

Focus on Forms (FoFS) is an approach to instruction directed at teaching preselected linguistic items in activities where the learners' primary focus of attention is on linguistic form, rather than on the meaning being conveyed (Long, 1996, 1997).

Focus on Meaning (FoM) is an approach to instruction directed at engaging learners in comprehending and producing messages in the target language where the learners' focus of attention is exclusively on meaning (Long, 1996, 1997; Long & Robinson, 1998).

Focused communication task or Focused communicative task is an activity that meets the requisite characteristics of a task, has been designed to predispose learners to comprehending and producing specific target language features that currently are the focus of instruction (see Focused task), and involves learners in communication in the target language for the purpose of completing the assigned task goal (R. Ellis, 2002, 2003; Nobuyoshi & Ellis, 1993). In this study, a focused communication task is operationalized as a focused task that involves learners in oral interaction with other TL speakers.

Focused task is a task that, in addition to developing learners' overall ability to communicate in the target language, has been designed to induce their incidental attention to some specific linguistic features (e.g., grammatical structures) while processing input or output in the target language (R. Ellis, 2003; Fotos, 2002; Nassaji & Fotos, 2004). For example, the learners will be processing expressions of frequency (e.g., "once a week," "twice a month," "every Saturday," etc.) if their task is to create and administer a questionnaire about how often their classmates complete certain house chores (e.g., vacuuming, shopping for groceries, doing laundry, etc.).



Foreign language teaching is teaching a language other than the learners' L1 outside of the target culture, as opposed to second language teaching that takes place when the target language is taught within the target culture (Spada & Lightbown, 2008b).

Form is all the grammatical features of the language, both morphological and syntactic (Doughty & Long, 2006).

Meaningful interaction is interaction in the target language among learners or between learners and native speakers of the language aimed at sharing real-life information that is not known to the interlocutor versus answering so-called display questions (e.g., "What color is this book"? ) or otherwise demonstrating ability to use target language correctly to express meaning prompted by someone else such as in translation exercises, structural drills, and so forth. Meaningful interaction can include providing information, exchanging opinions, solving real-life problems, and so forth (Brown, 2001; R. Ellis, 2003).

Nonfocused task is a task that is not specifically designed to encourage use of any particular linguistic features but rather to develop general ability to communicate in the target language as opposed to a Focused task (R. Ellis, 2003). For example, if the group task is to compile a ranked list of the main challenges that Western businesses face in Russia, it gives the learners an opportunity to use any TL items they have acquired so far unless they are primed in some way in the pretask phase for the use of a specific structure.

Preemptive focus on form is focus on form that does not arise from a learner error or a communication breakdown; it can be the result of a learner-initiated inquiry or a planned intervention initiated by the teacher (R. Ellis, 2001; Long, 2000).

Reactive focus on form is focus on form that occurs as a result of learner errors in the linguistic code or learners' inability to express the intended meaning accurately and concisely (R. Ellis, 2001; Long, 2000).

Second language teaching is teaching the language of study, other than the learners' L1, within the target culture, as opposed to foreign language teaching that takes place outside the target culture (Spada & Lightbown, 2008b).

Standardized-mean-difference effect size is the effect size measure that is calculated by subtracting the mean of the control or comparison group from the mean of the experimental treatment group and dividing the difference by the pooled standard deviation (Cohen, 1977; Norris & Ortega, 2000).

Standardized-mean-gain effect size is the effect size measure that is calculated by subtracting the mean-pretest-score value from the mean-posttest-score value and dividing the difference by the pooled standard deviation of the pre- and posttest values (Norris & Ortega, 2000).

Target culture is the country or area where the language of study is spoken (Brown, 2001).

Target language is the language being studied, other than the learner's L1, in a second language setting (i.e., within the target culture) or a foreign language setting (i.e., outside the target culture; Brown, 2001).

Target structure is the specific morphological or syntactic grammatical form that currently is the focus of instruction (Fotos, 2002; Larsen-Freeman, 2001a, 2003).

Task is an instructional activity completed in the target language that is defined in the field of foreign and second language teaching as a real-world activity with a nonlinguistic purpose for the learners' interaction and a specified observable outcome (R. Ellis, 2003).

Task-based interaction is conversational interaction in the target language that takes place among nonnative learners of this language or between nonnative learners and native speakers (in pairs or small groups) while completing classroom tasks (Keck et al., 2006).

In this study, task-based interaction is the independent variable operationalized as oral verbal exchanges among task participants that occur in the process of completing assigned focused (structure-based) communication tasks (see Task, Criterial features of a task, Focused task, and Focused communication task).

Task-based language teaching (TBLT) or task-based instruction (TBI) is an approach to FL or L2 instruction within the communicative approach that utilizes classroom activities that meet the requisite characteristics of tasks (see Task and Criterial features of a task) as curricular units. TBLT involves learning TL through performing communicative functions through the use of TL (R. Ellis, 2003; Nunan, 1989; Skehan, 1998), for example, solving a real-world problem, formulating a joint plan of action, predicting the outcome of an event, and so forth.

Task-supported language teaching or Task-supported instruction is instruction that utilizes tasks in addition to other types of classroom activities including possibly more traditional ways of presenting and practicing specific linguistic features (R. Ellis, 2003). For example, a target structure may be introduced through direct teaching of the associated rule and examples of its usage before learners complete focused tasks involving this structure.

Test of homogeneity is a  $Q$  statistic used to evaluate the computed effect sizes across the included primary studies for homogeneity. Testing for homogeneity before estimating the mean effect size is carried out to learn whether the effect sizes share a common population effect size. When the effect sizes are not homogeneous, their mean is not considered to be a good estimate of the population mean, and the differences in effect sizes may be associated with different study characteristics treated as potential moderator variables (Lipsey & Wilson, 2001).

Within-group contrast is a contrast between the performance of the same group (typically, the experimental group) on the pretest and a posttest. It can be expressed by the standardized-mean-gain effect size (Lipsey & Wilson, 2001; see Standardized-mean-gain effect size).

### Summary

The issue of what constitutes effective grammar instruction frequently gives rise to heated debates in the field of foreign and second language teaching. This study employed a meta-analytic approach to examine research into the effectiveness of focused oral-communication tasks involving student interaction in the classroom as an instructional technique for improving mastery of specific grammatical features (i.e., target structures). The meta-analysis involved quasi-experimental and experimental studies where the treatment includes teaching of foreign or second language grammar through interactive classroom practice activities that by design predispose learners toward using specific targeted structures repeatedly, but, unlike mechanical drills, require the learners to engage in exchange of real meaning. The results of this study serve to inform teachers, curriculum designers, teacher educators, and supervisors about the effectiveness

of using tasks in the teaching of TL grammar. Comparing the effect sizes across studies allowed an examination of various moderator variables that influence the effectiveness of interaction that occurs during focused communication tasks. Therefore, within the limitations presented in chapter V, this study has helped inform best practices in the design and classroom implementation of focused oral-communication tasks in foreign and second language teaching of adult learners.

### Forecast of the Study

To give the readers a sense of organization, the study starts with an introductory chapter (present chapter) containing the background and need for the investigation of the relationship between task-based interaction and acquisition of grammatical structures that are the focus of instruction. Chapter II presents the review of the relevant literature. It provides the historical perspectives that help understand the differing positions on the effectiveness of teaching grammar through interaction. Chapter II also expands on the theoretical framework, presents the discussion of variables moderating the effectiveness of task-based interaction, and provides a detailed review of a previously completed meta-analysis in the domain (Keck et al., 2006) that is most closely related to the focus of the present investigation.

Chapter III is the *Methodology* section where the methodology of meta-analysis used in the present study is described. The research design, search procedures, inclusion and exclusion criteria, and data analysis procedures also are explained. The results of the investigation are reported in chapter IV. In line with the established meta-analytic tradition in the field of SLA, chapter IV consists of two main parts: (a) the *Research Synthesis* section that summarizes various characteristics of the primary studies included

in the meta-analysis (e.g., educational settings, learner characteristics, tasks of types used as instructional treatment, etc.) and (b) the *Quantitative Meta-Analytic Findings* section that presents the results of the meta-analysis by research question. Conclusions drawn from the investigation including limitations and implications of the study for future research are presented in chapter V.

## CHAPTER II

### REVIEW OF LITERATURE

The purpose of the current study was to investigate the effectiveness of task-based interaction (i.e., interaction that occurs in so-called focused communication tasks) in form-focused instruction (FFI; see Appendix A for a list of abbreviations used in this study), that is, in teaching target language (TL) grammar. In chapter I, brief argumentation for the use of this classroom instructional technique that involves teaching of grammar through meaningful interaction in the TL was provided, and the main theoretical frameworks associated with task-based interaction were introduced. These frameworks are task-based language teaching (TBLT) and the Focus on Form (FoF) approach. The limitations of the previous meta-analytic investigations were presented, and the need for a more focused and fine-tuned investigation of the effectiveness of task-based interaction in acquisition of grammatical items was outlined.

Chapter II builds upon the argument presented in the introductory chapter. The first six sections of chapter II further develop the theoretical rationale for the use of focused communication tasks in language teaching, whereas the remaining sections present a discussion of potential moderator variables, a discussion of dependent variables that are used typically in primary studies in the domain, and a review of the previous meta-analysis (i.e., Keck et al. [2006]) that is most closely related to the purpose of the present study.

In the first section titled *Historical Perspectives*, a brief overview is provided of so-called precommunicative approaches that were dominant prior to the 1980s and continue to have a strong influence on the way some classroom practitioners

conceptualize and teach grammar. The subsequent sections titled *Communicative Competence and Communicative Language Teaching*, *Role of Input and Output in Foreign and Second Language Teaching*, *Role of Interaction in Foreign and Second Language Learning*, and *Skill Acquisition in Foreign and Second Language* help to develop argumentation in favor of teaching grammar through interaction and focus on key relevant theoretical concepts as well as hypotheses about how foreign language (FL) and second language (L2) acquisition occurs in learners: (a) the concept of communicative competence, (b) the input hypothesis, (c) the noticing hypothesis, (d) the output hypothesis, (e) the interaction hypothesis, and (f) the skill acquisition theory as it applies to the development of ability to use grammatical items accurately and appropriately. The section that follows is titled *Task-Based Language Teaching*; it provides a more in-depth discussion of TBLT than was provided in chapter I, including the definition and criterial features of classroom tasks (vs. exercises and free, unstructured conversation) that were used in the present meta-analytic study in order to determine whether a particular treatment used in a primary study was indeed task-based in nature.

The sections and subsections that follow contain an overview of various task-related, learner-related, and contextual variables that can moderate the effects of task-based interaction. These sections are *Types of Tasks as Moderator Variables*, *Role of Individual Learner Differences in Task Performance*, and *Other Task-Related Moderator Variables*. This discussion of possible moderator variables provides the basis for coding categories for this meta-analysis that are presented in chapter III. The section titled *Pedagogical Grammar and Language Acquisition* explains how the FoF approach that is



central to the current understanding of pedagogical grammar is different from both Focus on Forms (FoFS) and Focus on Meaning (FoM) and serves to situate task-based interaction as an FoF instructional technique before introducing such additional moderator variables as the type of target structure and degree of its task-essentialness.

The *Measures of Acquisition of Target Grammatical Structures* section presents a discussion of the types of outcome measures used to assess acquisition of target grammatical structures and of related measurement issues. It provides the background for understanding the dependent variable(s) involved in primary studies included in the present meta-analysis. Chapter II concludes with a detailed discussion of the meta-analytic investigation of the empirical link between task-based interaction and L2 acquisition conducted by Keck, Iberri-Shea, Tracy-Ventura, and Wa-Mbaleka (2006) that is related most closely to the topic of the present meta-analytic study.

### Historical Perspectives

A historical overview of beliefs about effective teaching of grammar is important for understanding the conflicting positions and debates that surround this subject in the field of second language acquisition (SLA). Known limitations associated with purely communicative approaches to teaching language have led some practitioners and researchers to believe that a return to the precommunicative methodologies is necessary for the development of sufficient grammatical accuracy in learners. This misconception sometimes results in language teachers overlooking valuable opportunities to improve learners' mastery of grammar through communicative tasks.

Precommunicative language teaching methodologies equated learning of the TL with the study of its grammatical system. For example, the grammar-translation method

equated learning of the language with the analysis of its grammatical structure. The audiolingual method that followed it was rooted in the behaviorist learning theories and emphasized the learning of correct grammatical patterns through repetition and drills (Purpura, 2004). Grammar-translation emphasizes deductive methods, whereas audiolingualism is supposed to stimulate inductive learning. Nevertheless, the two methods have something in common, that is, they both separate teaching of grammatical form from communicative meaning, and it is difficult for learners to make connections between different parts of the grammatical system and to understand how these parts relate to each other (Nunan, 1999).

Both the grammar-translation method and audiolingualism share an assumption that SLA is a linear process, that is, that learners can master one TL item at a time to perfection. In reality, learners acquire many structures imperfectly at the same time and then restructure their understanding of the language in complex, nonlinear ways as they reach qualitatively new proficiency levels (Lightbown & Spada, 1993; Nunan, 1999). Moreover, the behavior of a particular linguistic item in a learner's *interlanguage* (i.e., the developing implicit system) is frequently unstable because its rate of accurate use can increase or decrease at different times for various reasons, including that of interaction with other newly learned items. A learner's interlanguage development is prone to temporary deteriorations, for example, backsliding (i.e., when the learner's accuracy of use of a particular TL item drops after initial success; Selinker, 1972) and U-shaped learning (i.e., when the accuracy of use of a particular item drops but then comes up again; Kellerman, 1985).

Therefore, Nunan (1999) rejected the so-called “building block” metaphor, that is, the idea that learners, through a systematic approach designed for them, will build a solid “wall” of language proficiency “brick by brick.” He favored a more organic metaphor of a “garden” where all “plants” (i.e., language items) grow within the same timeframe but not in the same way or at the same rate. Some items may slow down in their growth and even “wilt” temporarily due to different environmental factors and interaction with other items. This metaphor is not meant to suggest that teachers and instructional designers should not plan for any systematicity at all in how the learners will come in contact with new TL items, but it vividly demonstrates the multidirectional, multifaceted, and multicausal nature of language development in learners.

The view of SLA represented through this metaphor clearly is incompatible with precommunicative methods (i.e., grammar-translation method and audiolingualism). Starting in the late 1970s and early 1980s, the work of Krashen (1981, 1982), combined with the development of the concept of communicative language competence and CLT, has led some researchers and classroom practitioners to reject the need for instructed grammar (Byrd, 2005). Based on a drastically defined distinction between conscious language learning and unconscious acquisition processes, Krashen (1981, 1982, 1993) claimed that conscious learning, including conscious learning of grammatical form, never leads to true language acquisition. Thus, using the pendulum analogy that is common in the SLA field, one can say that the pendulum swung from equating the teaching of language with teaching a set of explicit rules to the so-called *noninterface* position (Krashen, 1981, 1982) that did not recognize a link between conscious learning and developing real, functional ability to communicate in the language.

Due to the documented lack of learner success in acquiring target-like language forms through teaching methodologies grounded in the noninterface position (Higgs & Clifford, 1982; Swain, 1985), the SLA field moved in the direction of giving heightened attention to the teaching of grammar. For some classroom practitioners, this trend represented a return to the so-called traditional approaches, that is, to explicit explanations of rules followed by obligatory practice in nonauthentic, teacher-created, decontextualized, sentence-level drills (Cobb & Lovick, 2008). Thus it is reported that some classroom practitioners believe that the pendulum has now swung back from purely meaning-based communicative teaching to the traditional approach (Larsen-Freeman, 2001b; Macaro, 2003). Admittedly, a certain backswing of the “grammar pendulum” has occurred as a reaction to the lack of development of grammatical accuracy in students’ speech in purely meaning-based classrooms where attention to form was absent. According to Byrd (2005), however, the SLA field has managed to “correct the course to less extreme positions” (p. 551).

In fact, rightful rejection of the noninterface position does not at all signify the return to exclusive reliance on precommunicative techniques. Between purely meaning-based activities and old-fashioned drills lies a wide range of meaningful activities that include attention to language form. The question became, therefore, not *whether* to teach grammar but rather *how* to teach it. The possibilities include both receptive activities that assist learners in figuring out and mapping the form-meaning connections (Lee & VanPatten, 2003; VanPatten, 1996) and meaningful productive activities in which learners do not regurgitate meaning created by someone else but express their own

newly-created meaning in utterances generated for a true communicative purpose (R. Ellis, 2003; Larsen-Freeman, 2001a, 2001b, 2003; Nassaji, 1999).

In the contemporary view, the many facets of targeted grammar teaching include diverse task-based and text-based activities that may utilize printed passages, audio, video, pictures, and real objects in the classroom (Cobb & Lovick, 2008). In fact, any deductive or inductive learning activity that focuses the learners' attention on form, including discourse-based and content-based activities, constitutes formal grammar instruction (Celce-Murcia, 1992; R. Ellis, 2003). Balanced grammar teaching can include both implicit (i.e., indirect) and explicit (i.e., overt, direct) techniques. In contrast to Krashen's (1982) noninterface position that asserted that explicit and implicit knowledge are two separately organized knowledge systems, other researchers believe that explicit knowledge can facilitate the development of implicit knowledge. For example, DeKeyser (1998), who advocated the *strong interface* position, argued that explicit knowledge converts to implicit knowledge when automatization of explicit knowledge takes place. R. Ellis (1994a) who supported the *weak interface* position believed that explicit knowledge at a minimum indirectly facilitates implicit knowledge.

One of Krashen's (1982) assertions that have survived empirical testing is that there appears to be a natural order of acquisition for TL morphemes and structures. The *morpheme order studies*, the purpose of which was to determine whether the natural order of acquisition could be overturned by instruction, have provided evidence in support of this claim (Nunan, 1999). These results were disappointing for those who were in favor of making strong claims about the relationship between instruction and acquisition because not a single study showed that the order that is followed by the

learners' interlanguage could be changed by instruction. According to Lee and VanPatten (2003), research findings revealed that "natural learning processes always assert themselves over outside intervention" (p. 116).

Pienemann (1984, 1989) proposed that there are psychological constraints that govern whether attempts to teach specific target forms to learners will be effective. Formal instruction will succeed only if the learners have reached a developmental stage where they are psychologically and cognitively ready for a specific TL structure (Pienemann, 1984). The implication is that no amount of quality instruction will result in true acquisition of a TL structure for which the learner is not ready developmentally. Nevertheless, communicative classrooms that integrate formal instruction and opportunities for interaction were shown consistently to be superior to traditional classrooms and also to immersion (i.e., input-rich programs without formal language instruction; Spada, 1990).

Although there is a wide variety of possible classroom techniques, actual teaching of grammar seems to gravitate toward one of the two extremes: (a) continued use of teacher-fronted presentations and drills and (b) complete rejection of the teaching of grammar (Byrd, 2005). Unfortunately, some of the most frequently overlooked grammar teaching techniques ultimately may be the most beneficial ones for developing grammatical accuracy in FL learners, for example, the focused communication tasks that are investigated in this study. The use of such tasks supports the development of the learners' communicative competence that is the core of the communicative approach to teaching FL and L2.

## Communicative Competence and Communicative Language Teaching (CLT)

Savignon (2001) described her model of communicative competence, first proposed in 1972, as consisting of four major interrelated components: *linguistic competence* (i.e., ability to code messages according to TL norms that sometimes is referred to as *grammatical competence*), *discourse competence* (i.e., understanding TL text organization and organizing one's own textual output appropriately), *sociocultural*, or so-called *sociolinguistic, competence* (i.e., understanding of cultural values and norms underlying meaning), and *strategic competence* (i.e., ability to plan and execute TL interactions effectively notwithstanding limitations in language mastery). These components cannot be developed or measured in isolation, and an increase in one interacts with the other components, producing an increase in the overall communicative competence (Canale & Swain, 1980).

Savignon (2001) clarified that linguistic, or grammatical, competence neither is based on any single theory of grammar nor includes ability to state grammar rules. Instead, a learner can demonstrate grammatical competence by using a rule for appropriate interpretation, expression, or negotiation of meaning. Larsen-Freeman (2001b) argued that grammar not only involves intrasentential patterning but also frequently deals with the interconnectedness of utterances as well.

Communicative competence entails both TL *fluency* and *accuracy*, even though these two major determinants of language proficiency frequently are perceived in terms of a dichotomy and almost in opposition to each other. *Accuracy* typically is understood to refer to grammatical accuracy, however, lexical accuracy, spelling, and pronunciation also can be included in this aspect of language proficiency. *Fluency* entails ability to

understand TL with relative ease and to participate in spontaneous, real-time TL communication (Byrd, 2005). Byrd warned, however, that, as is frequently the case with dichotomies, this particular one can lead to false distinctions. The two notions really are not mutually exclusive: fluency requires a certain degree of accuracy for both comprehending the interlocutor's utterances and for creating one's own comprehensible contributions to communication, whereas accuracy without fluency would most likely result in labored production that could not measure up to any real-life functions that need to be performed in real time. Task-based interaction that occurs in focused communication tasks and is the focus of the present meta-analysis serves the purpose of developing both accuracy and fluency in language learners.

#### Role of Input and Output in Foreign and Second Language Learning

This section presents a discussion of the role that TL input as well as TL output produced by learners play in the learners' interlanguage development. The need to elicit meaningful learner output in the TL in addition to providing rich, authentic TL input is what underlies the rationale for using communicative classroom tasks that require TL production, rather than merely comprehension, on behalf of the learner.

Krashen's (1985) input hypothesis posited that learners progress in acquiring the TL when they receive messages that are linguistically one step beyond their current stage of development. He asserted that input-based language development takes place as long as the input is comprehensible, that is, at the " $i+1$ " level where  $i$  represents the learner's current state of language development and  $+1$  represents a hypothetical increment within the learner's reach. If comprehensible input is provided and the learner's *affective filter* (i.e., internal screen of emotion such as anxiety, fear of embarrassment, etc. that can



block acquisition) is down, TL acquisition will take place (Krashen, 1985). Just as Krashen, many researchers recognize the centrality of rich, authentic input of the natural language in FL and L2 teaching; however, most believe that comprehensible input is a necessary but not a sufficient condition for language acquisition (Lightbown & Spada, 1993; Long, 1996).

Schmidt (1983, 1990) pointed out that the first prerequisite for acquisition of a language item is the learners' noticing of this item in the input (Chaudron, 1985; Sharwood-Smith, 1993; Van Patten & Cadierno, 1993). In opposition to Krashen's notion of purely subconscious acquisition, Schmidt, who carried out a study of his own experiences of studying Portuguese in Brazil, found out that he only acquired items that he had noticed and attended to consciously. This finding led the researcher to formulate the so-called *noticing hypothesis* (Schmidt, 1983, 1990; Schmidt & Frota, 1986). Thus, Krashen's (1981, 1982, 1985, 1993) claims about sufficiency of comprehensible input were refuted, and researchers became concerned with what makes input comprehensible and thus makes it possible for comprehended input to become *intake*, that is, input that has been filtered and processed by the learner (Schmidt, 1990). In other words, intake represents that subset of the linguistic data in the input that learners attend to and hold in working memory during real-time comprehension.

Regarding the learners' noticing of grammatical items, empirical research findings have provided evidence in support of the effectiveness of such implicit techniques as *textual enhancement*, that is, highlighting the target structure in the text through change of color or use of italics, bold-face fonts, capital letters, underlining, and so on (Jourdenais, Ota, Stauffer, Boyson, & Doughty, 1995), and *input flooding*, that is,

choosing texts in which a particular structure abounds (Wong, 2005). An example of an explicit technique that promotes noticing is a so-called *consciousness-raising task*, that is, a task that requires learners to make a generalization about how the target structure functions in the TL (R. Ellis, 2003; Fotos, 1994; Larsen-Freeman, 2001b; Pica, 2009). After the target structure is noticed and processed, this processed input can be converted to *uptake*, that is, learner growth through internalization and incorporation of the TL feature into the interlanguage. Uptake that occurs during classroom TL interaction is defined more narrowly by Lyster and Ranta (1997) as a learner's utterance produced in reaction to the interlocutor drawing the learner's attention to some aspect of the learner's previous utterance with an intent to make it more target-like (i.e., correct or appropriate).

One of the most beneficial techniques in promoting the appropriate interpretation of grammatical form by learners is the so-called *input processing*, or *processing instruction*, proposed by VanPatten and his associates (Cadierno, 1995; Lee & VanPatten, 2003; VanPatten, 1993, 1996; VanPatten & Cadierno, 1993). Input processing activities push learners to attend to properties of a language structure in receptive (vs. productive) activities and to connect different variations of this structure with differences in meaning. VanPatten (1996) pointed out that traditional classroom instruction typically moves from presentation of a new grammatical structure directly into production activities ranging from mechanical drills to more meaningful communicative practice. As a result, the learners are not given ample opportunity to process the new input and to foster the necessary form-meaning connections needed to convert the input to intake. Input processing, however, stimulates careful form-meaning mapping for new TL features before learners are encouraged to produce them.

VanPatten and Oikarinen (1996) compared the effects of explicit instruction (i.e., instruction that involves overt presentation of grammar rules) plus input-based processing activities (i.e., processing instruction) with the effects of explicit instruction only and input-based processing activities only for a group of US high-school students studying Spanish. The researchers reported that the gains demonstrated by the explicit instruction only group were not as large as those of the other two groups. Thus VanPatten and Oikarinen concluded that input processing instructional activities are more beneficial than traditional explicit instruction.

Even though processing instruction is recognized to be quite useful in developing grammatical competence, it does not develop the learner's ability to use grammar to speak (Lee & VanPatten, 2003). The latter purpose is served by a variety of structured output activities that promote both fluency and accuracy and combine attention to form with attention to meaning. In other words, Lee and VanPatten (2003) promoted "one kind of instruction for developing the underlying system and another for tapping that system and promoting the development of fluency" (p. 3). Classroom teachers intuitively agree that the loop is incomplete without learner output, or language production.

Lee and VanPatten's (2003) assertion is in line with Swain's (1985, 1993) *output hypothesis* that posited that learners modify their output to get the meaning across when they are forced to do so by negative feedback (i.e., the interlocutor signaling lack of understanding of the intended meaning). Swain (1985) reported that the results obtained by sixth-grade children in a French-immersion setting in Canada were substantially lower than the results obtained by their native French-speaking peers on a number of grammatical, discourse-related, and sociolinguistic language-acquisition measures, even

though the children in the immersion setting had plenty of access to rich comprehensible input in French. Swain speculated that the lack of French proficiency, including grammatical competence, in nonnative-speaking (NNS) children participating in the immersion program where they primarily engaged in comprehension activities was a result of lack of opportunities for production of output in the TL.

In the same year, Montgomery and Eisenstein (1985) compared the learning outcomes for grammar for English as a Second Language (ESL) community college students who participated in a grammar course with an added oral communication component with the outcomes for the control group who only took the grammar course. The study findings suggested that formal grammar instruction supplemented with opportunities to communicate using the newly learned structures led to greater improvements in not only fluency but also grammatical accuracy than grammar instruction alone. Other research findings confirmed that instruction and opportunities to communicate in the TL out of class were both necessary, and learning of grammar occurred when learners had an opportunity to “notice the gap” between their own production and target forms in output activities (Schmidt & Frota, 1986). In comprehension, learners have been shown to rely on semantic, rather than syntactic, processing of input as well as on contextual clues and their own general world knowledge when trying to understand the meaning of input (Sharwood-Smith, 1981; White 1987). Therefore, Swain (1993) as well as Kowal and Swain (1994) argued that so-called *pushed output* is needed in order to force learners to switch to syntactic processing from primarily semantic processing of the input.

Additionally, without output there would be absolutely no way to measure and assess learning. Learner-produced TL output also allows for the development of automaticity in the use of acquired linguistic resources (Swain, 1995). Automaticity is not to be understood as simple habit formation achieved via repetition and drills (as in traditional grammar instruction) but rather as a shift from controlled to automatic processing by means of repeated activation of language processing, according to McLaughlin's (1987) information processing model. Automatic processing (vs. controlled processing) is characterized by easy and swift retrieval of knowledge and does not require engaging the learner's focal attention (DeKeyser, 1997, 1998, 2001, 2007; Segalowitz, 2003). Automatization of certain aspects of performance means that the students' limited attentional resources are freed to be used for other purposes (DeKeyser, 2007; Larsen-Freeman, 2001b). In addition to serving the purposes of automatization, output allows for natural hypothesis testing about the linguistic patterns that the learner attempts to use because it generates feedback or response from the interlocutor (Swain, 1995).

In summary, although purely input-based processing activities can be very beneficial in mapping and internalizing form-meaning connections, output-based activities are paramount for language learning. Output production affords learners a natural way to test their hypotheses about how specific language forms function and to receive so-called negative feedback (i.e., error correction or evidence that their message has not been understood). Additionally, production of TL output forces learners to process language syntactically as well as semantically and thus facilitates both the development of grammatical knowledge and automaticity that is discussed in more detail

in the section titled *Skill Acquisition in Foreign and Second Language*. The need for output helps build the theoretical foundation for the use of structure-based communication tasks that are the purpose of the present meta-analysis. Because the majority of these classroom tasks require the interlocutor to respond to the produced output, the rationale for using such tasks is developed further in the next section.

### Role of Interaction in Foreign and Second Language Learning

The discussion of the role of interaction provided in this section builds the foundation for understanding the need for classroom activities that require learners to interact with others in the TL in addition to purely input-based or purely output-based activities. This section provides an overview of specific processes that occur during interaction that are beneficial for FL and L2 learners' interlanguage development.

Generation of truly meaningful output is only possible in the process of interacting with an interlocutor. Long's (1981, 1996) *interaction hypothesis* posited that comprehensible input that is interactionally modified promotes language acquisition. In this approach, conversation between a nonnative speaker (NNS) and a native speaker (NS), or among NNSs, is not considered merely a stage for practicing the previously learned language features but rather as a means for learning the language (Gass, 1997). Modified interaction happens when communication is repaired after an initial mismatch between the speaker's intentions and the listener's interpretation of the message resulted in a complete or partial miscommunication. It usually is repaired by means of clarification requests, use of "or"-choice questions, comprehension checks, confirmation checks, clarification requests, rephrasing and elaboration of the original message, and so forth (Long, 1996). This process of improving message comprehensibility (Pica, 1994) is

called *negotiation for meaning* in language learning; however, interlocutors can negotiate either for *meaning* or *form* or for both meaning and form simultaneously because these negotiation processes frequently are intertwined. The need for negotiation for meaning typically does not arise when interlocutors are engaged in asking and answering questions in the TL about content that is already known to both of them (i.e., so-called display questions). Negotiation occurs naturally, however, when interlocutors are involved in information-gap, problem-solving, or consensus-reaching tasks described later in this chapter.

A few studies linked conversational adjustments with improved comprehension that is then believed to lead to acquisition. For example, Pica, Young, and Doughty (1987) compared the comprehension of 16 NNSs enrolled in preacademic college ESL courses on a task presented by an NS under two conditions: (a) premodified input (i.e., input that had been simplified for the learners beforehand) and (b) interactionally-modified input (i.e., input that was made comprehensible to the learners as a result of negotiation of meaning in the TL between the learners and their NS interlocutor). The task required NNSs to listen to the NS give directions for selecting and placing 15 objects on a board depicting an outdoor scene. The results showed that learners demonstrated better comprehension under the interactionally-modified-input condition in which the NS engaged in repetition and rephrasing of original input based on the reactions received from the NNSs. The result of a *t* test showed that the difference between the higher scores for correct selection and placement of the objects for the interactionally-modified-input group and those for the premodified-input group was statistically significant.

In a follow-up study, Pica (1991) compared three conditions: (a) negotiation (i.e., active involvement in interaction with others), (b) observation (i.e., observing the negotiation conducted by others without actively participating), and (c) listening (i.e., receiving premodified input that already included redundancy features such as repetitions and paraphrasing that typically are present in negotiated input). Pica reported that the participants under the negotiation condition demonstrated better comprehension than those under the other two conditions. The researcher, therefore, suggested that redundancy features that are generated as a result of conversational interaction, rather than those provided upfront, may lead to better comprehension.

Pica, Holliday, Lewis, and Morgenthaler (1989) reported that NNS learners themselves modified their output when their NS interlocutors asked for clarification or otherwise indicated difficulty in comprehending utterances produced by NNSs. In addition, these researchers provided evidence that the modifications made by the learners were related to morphology rather than semantics. Such negotiation of grammatical form is believed to facilitate TL development (Lyster & Ranta, 1997).

An additional argument in support of the role of interaction comes from the assertion that social interaction facilitates any learning process. Van Lier (1996) clarified Vygotsky's (1986) position and emphasized the centrality of social interaction in the pedagogical process, that is, the view that the learners' different perspectives, knowledge, and strategies create a cognitive conflict that forces them to reflect on and question their understanding. In the process of resolving this conflict, new perspectives, knowledge, and strategies are created. According to van Lier (1996), learners construct new language knowledge through socially mediated interaction. In general, group work is considered to



be an essential element of CLT that offers important benefits (Brown, 2001; Brumfit, 1984) such as a more positive affective climate, greater practice opportunities (i.e., greater amount of TL output produced by each student during class hour), and so forth. As long as the learners' group activities are well-designed and monitored by the teacher, the use of group work is beneficial to language acquisition.

Numerous empirical research studies with an interactional focus have been published up to date, including in the 2000s. For the most part, their findings reemphasized the benefits of interaction to FL and L2 development (Mackey, 2007). These studies focused on the effectiveness of negotiation (de la Fuente, 2002), interactional feedback (Mackey, 2006; Mackey & Oliver, 2002), interactional modifications (Pica, Kang & Sauro, 2006), learners' perceptions about interactional processes (Gass & Lewis, 2007; Mackey, 2002), and so forth. A few of these studies investigated various aspects of acquisition of specific TL grammatical items in a variety of educational contexts of FL and L2 learning in both child and adult learners (Ayoun, 2001; Iwashita, 2003; McDonough, 2006; Pica, Kang, & Sauro, 2006). As stated in chapter I of this study, two meta-analyses of effectiveness of interaction in TL development have been published (Keck et al., 2006; Mackey & Goo, 2007). These meta-analyses reaffirmed general effectiveness of interaction through rigorous statistical procedures that compared effect sizes for treatments involving interaction with other types of treatments across eligible primary studies. Focused (structure-based) communication tasks that predispose learners to using specific target grammatical structures are discussed in more detail in the subsections titled *Focused and Nonfocused Tasks* and *Pedagogical Grammar and Language Acquisition*.

Table 1 summarizes the relevant SLA hypotheses that have been discussed in this chapter for the purposes of building the theoretical rationale for the use of TL focused (structure-based) communication tasks that are the focus of investigation in the present study. The next section further develops the rationale for the use of focused communication tasks in the classroom from the point of view of skill-acquisition theory that is applicable to learning all complex cognitive skills (DeKeyser, 2001, 2007; Leeman, 2007). Later in this chapter, focused communication tasks are positioned as one of the instructional techniques under the FoF approach to teaching grammar that is one of the main methodological principles of TBLT (Doughty & Varela, 1998; Doughty & Williams, 1998; Long, 1996; Long & Robinson, 1998).

Table 1

## Relevant Second Language Acquisition (SLA) Hypotheses

Hypothesis	Formulation	Pedagogical Implications
1. Comprehensible input (Krashen, 1982, 1985)	Learners acquire TL when provided with TL input slightly above their current level.	Authentic input at the appropriate level in and of itself is sufficient for TL acquisition.
2. Noticing (Schmidt, 1983, 1990)	Learners can acquire only those TL features in the input that have been noticed and consciously attended to.	Teaching techniques and activities that encourage noticing and processing of target structures in the input are necessary.
3. Output (Swain, 1985, 1993)	Learner-produced TL output, in addition to TL input, is required for successful acquisition.	Teachers need to set up activities that elicit learner-produced output in the TL.
4. Interaction (Long, 1981, 1996)	Modified interaction in the TL facilitates acquisition.	Teachers need to set up activities that involve learners in meaningful interaction in the TL.

### Skill Acquisition in Foreign and Second Language

Another important area of SLA research that positions L2 learning in line with acquisition of other complex cognitive skills is *skill-acquisition theory* (DeKeyser, 2001, 2007). The skill-acquisition approach is valuable because it helps build theoretical support for the need of meaningful L2 practice and reflects the integration of cognitive psychology and SLA theory (Leeman, 2007).

The building of the learner's interlanguage system does not equate merely with habit formation in the behavioristic sense (Macaro, 2003) but rather with constant, complex restructuring of knowledge representation in the mind of the learner (Lee & VanPatten, 2003). The true essence of TL acquisition is the system change that is believed to occur somewhere between input processing (i.e., processing of the form-meaning connections present in a new TL structure) and subsequent output processing (i.e., learners beginning to formulate their own previously unrehearsed utterances). This system change, or restructuring of the learners' interlanguage, according to Lee and Van Patten, involves two subprocesses: (a) accommodation (i.e., incorporation of a grammatical form into an implicit system) and (b) restructuring (i.e., the overall change in the knowledge of other forms as a result of this incorporation).

The nature of a learner's linguistic knowledge changes over the course of acquisition in such a way that it gradually becomes more available for use in communicative settings. On the one hand, Bialystock (1988, 1994a, 1994b) argued that linguistic knowledge starts out as *implicit* (i.e., unanalyzed, subconscious) knowledge and becomes more *explicit* (i.e., conscious) as the learner becomes more proficient so that it can be applied consciously in novel TL situations. On the other hand, in contrast to

Bialystock's position, Sharwood-Smith (1988) and DeKeyser (1998, 2007), among others, argued that the development of L2 proficiency is a process of automatizing explicit knowledge so that it eventually becomes implicit.

According to a widely accepted model in cognitive psychology, skill acquisition proceeds through three stages: (a) acquisition of declarative knowledge, (b) proceduralization, and (c) automatization (Anderson, 1983, 1993). Declarative knowledge of a TL grammatical structure may be understood as consisting of three dimensions: its form, its meaning, and the appropriateness of its use (Larsen-Freeman, 2001b) in the form of associated rules or examples (Segalowitz, 2003). Although it may be possible for a learner to utilize declarative knowledge in skill performance, the cognitive demands (i.e., memory and processing requirements) of relying on this factual knowledge about the target structure in the absence of a proceduralized skill are very high (Leeman, 2007). Advancement to procedural knowledge, that is, the skill of applying the rule in both receptive and productive processes involving the TL (Segalowitz, 2003; Zhou, 1991), results in lowering the cognitive load. Therefore, procedures that have relied on declarative knowledge of the target grammatical structure can now be combined with other procedures unrelated to the target structure, thus allowing the learner to attend simultaneously to several other aspects of the TL utterance in a more efficient manner (Larsen-Freeman, 2001b; Leeman, 2007; Skehan, 1998). The next stage is automatization of the procedural knowledge, at which point explicit knowledge of the target structure becomes virtually implicit (DeKeyser, 1997, 1998, 2001, 2007; Sharwood-Smith, 1988). Automatic processing, as opposed to controlled processing, allows for effortless comprehension and production of the target structure in the context of natural TL

interaction while the learner's attention is on meaning rather than on form (DeKeyser, 1997, 1998, 2001, 2007; Segalowitz, 2003). The true measure of successful acquisition of a target structure is the learner's demonstrated ability for spontaneous, relatively effortless, and errorless processing of the form as it comes up in communication, rather than ability to produce the form when prompted by a teacher or to provide the associated rule (Jackson, 2008; Nunan, 1999).

The concept of automaticity has important implications for FL and L2 pedagogy. DeKeyser (2007) argued that it truly takes an enormous amount of *deliberate* (i.e., intentional) and *specific* practice to become a skilled, advanced FL or L2 speaker. The question is what can be considered meaningful practice in acquisition of TL grammar because, according to Lightbown (2007) and DeKeyser (2007), not all kinds of TL practice bring desired results. In the behavioristic approach to language learning, the notion of practice is, for the most part, associated with drill-type, habit-forming, repetitive activities such as decontextualized structural drills (Larsen-Freeman, 2001b). For example, in a classic substitution drill for the English present progressive tense, if the teacher says, "I am drinking milk" and then prompts the learner to use "to wash my clothes" in a similar utterance, it is possible for the learner to respond, "I am washing my clothes" without understanding what the utterance means. Some drills arguably are more communicative in nature and, unlike purely mechanical drills, cannot be completed without the student fully understanding what is being said (DeKeyser, 2007). For example, if the learners are asked to answer the question, "Is the boy drinking milk or washing his clothes?" based on a picture, the response still will be controlled, but the learners definitely have to understand what they are saying (Macaro, 2003).

Although most researchers and classroom teachers attach a certain, even though a limited, value to drills (Lightbown, 2007), decontextualized structural drills do not go far enough in equipping learners with the ability to communicate because they do not represent *transfer-appropriate processing* (DeKeyser, 2007; Lightbown, 2007).

Lightbown explained that transfer-appropriate processing takes place when the initial encoding of information happens under the same conditions under which this information will be retrieved later. In other words, retrieval will be most successful when the processes that are involved during encoding are the same processes that are active during retrieval. For this reason, the activities of filling in the blanks with correct grammatical endings or completing a substitution drill (e.g., “I am drinking milk” – “to wash my clothes” – “I am washing my clothes”) do not represent transfer-appropriate processing if the learner’s goal is using grammar correctly in real communicative situations.

In addition to reproducing language models provided by others, learners need opportunities for creative language use (Nunan, 1999). Nunan explained that by creative use he does not mean having learners “write poetry” in class but rather having them complete activities that require recombination of learned language elements into new, previously unrehearsed utterances. Learners need to be given structured opportunities to use the language that they have been practicing in new and unexpected ways to achieve various communicative goals.

In cognitive psychology, Newell and Rosenbloom (1981) limited the definition of practice only to that part of learning that deals with improving performance on a task that the learner can already complete successfully. In FL and L2 learning, the purpose of practice is to decrease the time needed to complete the task, that is, to increase TL

fluency, and to reduce the error rate, that is, to improve grammatical accuracy (DeKeyser, 2007; Segalowitz, 2007). DeKeyser (2007) argued that practice is skill-specific, which suggests that success in appropriate structuring of TL utterances does not develop in reading and listening activities necessarily and that learners need targeted grammar-focused output practice. Considering the limited nature of the learners' attentional resources during language-task completion (Skehan, 1998), it is important to create conditions where deliberate grammar practice is not overshadowed by other processing and interactional demands of the classroom task such as finding precise and appropriate vocabulary, planning the interaction, organizing one's thoughts logically, observing politeness and other pragmatic norms, and so forth. All these considerations point to a necessity of controlled and tight-focused, yet meaningful, classroom practice.

The need for communicative practice of grammatical structures was underscored by Larsen-Freeman (2001b), who proposed teaching the skill of "grammaring," a term she coined for this purpose, as opposed to traditional teaching of grammar based on the knowledge-transmission model. Although the term "grammaring" does not appear to have taken root in SLA literature, the concept of improving learners' mastery of grammatical structures through structure-based (i.e., *focused*) communication tasks that is the focus of this study has received a considerable amount of support.

The skill-acquisition model presented earlier in this section views language learning as an increasing degree of implicitness of TL knowledge (DeKeyser, 2007). DeKeyser, who perhaps is one of the strongest proponents of the skill-acquisition theory in SLA, believed that adults (vs. children) initially rely exclusively on explicit processing in their comprehension of TL structures. Other researchers reported findings suggesting

that learners are able to process certain aspects of the TL syntax implicitly even at early stages of language development (Tokowicz & MacWhinney, 2005). Robinson (2005) conducted an empirical replication study investigating students' learning of grammar of an artificial language and the Samoan language under three conditions: (a) explicit, (b) implicit, and (c) incidental. The participants were 54 undergraduate students at Aoyama Gakuin University in Tokyo, aged 19 to 24 years, who were experienced FL learners. Robinson reported that the test of the variance of scores showed a statistically significant difference between learning outcomes under the explicit and implicit conditions  $F(36, 36) = .502$ , with the variance in implicit learning being statistically significantly smaller than the variance in explicit learning. There were no statistically significant differences between the variance in implicit and incidental learning or explicit and incidental learning. Robinson also investigated the relationship between learning outcomes and certain cognitive characteristics of the learners. He reported a statistically significant negative correlation between the learners' IQ and implicit language learning:  $r = -.34$ . The relationship is a weak one.

In general, research findings have provided evidence that, contrary to Krashen's (1982) contention, even though explicit and implicit knowledge are dissociable, they interact with each other (R. Ellis, 2006b). The extent to which form-focused instruction (FFI) contributes to the acquisition of implicit knowledge still remains a controversial and unresolved issue in SLA (Ellis, 2002). There are, however, studies that have provided evidence supporting the assumption that communicative practice, especially TBLT, can lead to interlanguage development. For example, in order to test an assumption that TBLT contributes to development of automaticity in language learners, De Ridder,



Vangenhuchten, and Gomez (2007) conducted an empirical research study that involved 68 intermediate-level students of Spanish in their early 20s at the Antwerp University in Belgium. The comparison group (35 participants) attended a traditional communicative course, whereas the experimental group (33 participants) attended a course that had a task-based component built into it. The researchers reported that the experimental (i.e., task-based instruction) group outperformed the comparison (i.e., nontask-based instruction) group on measures of automaticity. The results were  $t(66) = 6.06$ ,  $\eta^2 = .36$ , which is a large effect, for the criterion of the use of grammatical structures covered in the course;  $t(66) = 5.51$ ,  $\eta^2 = .32$ , which is a large effect, for vocabulary; and  $t(66) = 5.52$ ,  $\eta^2 = .32$ , which is a large effect, for sociolinguistic accuracy. The comparison group outperformed the experimental group on measures of pronunciation and fluency but no statistical significance could be established for fluency. The researchers speculated that the higher results achieved by the comparison group on measures of pronunciation and intonation could be explained by the fact that the comparison group participants had spent more time interacting directly with the teachers as compared with the experimental group participants who interacted with each other while performing tasks.

In summary, it appears that both explicit and implicit learning contribute to interlanguage development in adult FL and L2 learners. In any case, the role of skill-specific, transfer-appropriate TL practice (i.e., practice that promotes development of skills that are transferrable to situations of real communicative language use) in FL and L2 development cannot be overestimated. Arguably, TBLT provides learners with opportunities for transfer-appropriate processing of various language items that is needed

for effective skill acquisition. In particular, focused communication tasks that are the focus of investigation in the present meta-analysis facilitate the use of learned grammatical structures in new, unexpected ways that fit the communicative demands of specific real-life situations. The following section provides a detailed overview of TBLT.

### Task-Based Language Teaching (TBLT)

This section defines the role of TBLT within CLT and discusses such issues as the definition of a language task and its criterial features that distinguish it from nontask activities such as language exercises and free, unstructured conversation in the TL. The issue of the lack of a consensus as to what constitutes a task in the SLA field is discussed, and the operationalization of a communication TL task for the purposes of the present research study is presented. The section concludes with a summary of benefits and limitations of TBLT as an instructional approach.

CLT emphasizes development of communication skills and views communication in the TL not only as the goal but also as the means of TL development (Canale & Swain, 1980; Lee & VanPatten, 2003; Savignon, 1972; Widdowson, 1978). Consequently, there is an emphasis on classroom interaction among learners through a variety of games, role plays and simulations, information-sharing and problem-solving activities, and so forth (Savignon, 1972, 1983).

In the literature on FL and L2 teaching methodology, classroom activities typically are classified into so-called *tasks* and *nontasks* (R. Ellis, 2003; Nunan, 1989, 2004; Willis, 2004). Compared with its common usage in the English language, the term *task* has taken on specific meanings in SLA (Lightbown, 2007; Littlewood, 2004; Nunan, 2006), even though there is no consistency in the way this term is used in both research

publications and pedagogic literature (R. Ellis, 2003). The main characteristic of classroom tasks is that they enable students to learn TL by experiencing how it is used in real communicative situations (R. Ellis, 2003).

Nontasks mainly are two types of classroom activities: (a) exercises (e.g., drills that involve manipulation of language form but not manipulation of information, as well as language display activities such as answering comprehension questions about a passage) and (b) free (i.e., unstructured) conversation that involves a free exchange of ideas between interlocutors without any workplan (i.e., procedure to follow) or observable outcome (R. Ellis, 2003). As opposed to exercises and free conversational exchanges, classroom tasks are increasingly complex approximations of target tasks that the learners eventually will be expected to perform in the real world using the TL (Long, 1996; Long & Norris, 2000). The theoretical rationale for use of classroom tasks is found in Long's (1996) interaction hypothesis that postulated that interaction in the TL contributes to TL acquisition. Skehan (1998), Robinson (2001a, 2001b), and Van den Branden (2006), among other researchers, pointed out that classroom tasks give rise to a number of interactional and cognitive processes believed to enhance language acquisition.

TBLT, a development within CLT that has gained prominence since the 1980s, is an approach in which tasks, rather than texts, are considered to be primary curricular and instructional units (Long, 1996; Long & Crookes, 1993; Nunan, 1993). Task-based syllabi represent a more holistic approach to language teaching compared with traditional syllabi that are based on the notion that language should be broken into isolated linguistic units and presented to learners one unit at a time in a linear, cumulative fashion (Nunan,

1999). This latter type of approach to syllabus construction typically is referred to as *synthetic* due to the fact that learners are expected to integrate, or synthesize, the language items taught in this manner into a coherent functional system (Long & Robinson, 1998; Wilkins, 1976). A synthetic syllabus at any given time exposes the learner to limited samples of TL that incorporate only the language items that have been taught explicitly so far.

SLA research has pointed out numerous problems with synthetic approaches. First, actual TL development does not happen in small, predictable increments so that each new set of linguistic units can be mastered to perfection before a new set is introduced (Nunan, 1999). A learner's interlanguage development is prone to temporary deterioration defined as backsliding (Selinker, 1972) or U-shaped learning (Kellerman, 1985). Second, Pienemann (1989) proposed that there are psychological constraints that govern whether attempts to teach learners specific target forms will be effective. Formal instruction can be successful only if the learners have reached a developmental stage where they are psychologically and cognitively ready to acquire a specific TL structure (Pienemann, 1984). SLA research findings have demonstrated that FL and L2 learners naturally follow a certain order of acquisition of TL features, or so-called developmental sequences, that override the order in which these features are presented in textbooks (R. Ellis, 1994a; Kwon, 2005). For example, learners of English pass through set sequences in the development of negation and interrogatives (Pienemann, 1989; Schumann, 1979).

The recognition of these problems with synthetic syllabi has led SLA researchers to explore other types of syllabi. The so-called *analytic* approach to syllabus design exposes learners to chunks of TL as it occurs in the real world outside the classroom and

relies on the learners' ability to process and internalize TL features. The task-based approach to language teaching involves such an analytic syllabus where the selection of content and format for teaching activities is governed by real-life functions for which the learners will eventually use the TL (Long & Robinson, 1998; Wilkins, 1976). Littlewood (2004) pointed out that, unlike the well-known audiolingual method or the so-called direct method (i.e., teaching the TL without the use of the learners' first language [L1]) that have rather narrowly defined, identifiable characteristics, TBLT does not constitute a specific prescribed methodology but rather a flexible framework that can be used for a range of pedagogic purposes at different points in a teaching sequence.

Not all proponents of TBLT in the field of SLA agree on what exactly TBLT constitutes. For example, Long's (1985, 2000) task-based approach is in line with a stronger version of CLT that argues that language should be acquired only through communication (Howatt, 1984). Therefore, TBLT, as formulated by Long, treated a task as the principal, if not the sole, unit of the language curriculum and language assessment. In this approach, deliberate attention to language features occurs only as a result of a problem encountered by learners during task completion but is never planned or proactive (Long, 2000). As opposed to Long's approach, R. Ellis' (2003) conception of *task-supported*, rather than task-based, language teaching parallels the weaker version of TBLT that views tasks as a way of providing meaningful communicative practice for language items that may have been introduced in more traditional ways.

In summary, TBLT is consistent with the communicative approach to language teaching and appears to be aligned with the nonlinear nature of interlanguage development and the learners' internal constraints better than more traditional, synthetic

approaches to language teaching and curriculum design. SLA researchers' conceptualizations of TBLT vary from the strict approach that recognizes tasks as the only viable curricular units to more moderate versions that allow for integration of tasks with more traditional elements of language instruction (e.g., explicit rule explanations, measured use of drills and exercises, etc.). The present meta-analysis investigated the effectiveness of oral task-based interaction that occurs in FL and L2 classrooms while learners complete collaborative tasks.

### *Definition of Task*

Although tasks undoubtedly occupy a central place in SLA research as well as in language pedagogy (R. Ellis, 2003), the definitions of a task provided in the literature vary widely in terms of what their authors emphasize. This subsection provides a brief overview of some existing definitions and concludes with R. Ellis' definition of a communicative (or communication) TL task that was used in the present research study to determine whether the treatment described in each included primary study report can be considered to be TBLT.

According to Nunan (1989), a task is “a piece of classroom work which involves learners in comprehending, manipulating, producing, or interacting in the TL while their attention is principally focused on meaning rather than form” (p. 10). R. Ellis (2003) defined a task as a workplan and stressed that it requires learners to use TL pragmatically in order to achieve the desired propositional intent (i.e., to accomplish the needed communicative outcome such as to inform, justify, persuade, come to an agreement, etc.). As clarified by Samuda (2005), a good pedagogic task typically has some kind of data or content material as a starting point and requires learners to take some kind of action (e.g.,

processing or transforming) on these initial data as a means of reaching the given outcome.

A notably differing definition was offered by Long (1985) who did not emphasize the presence of language in tasks and simply defined a task as “a piece of work undertaken for oneself or for others” such as a multitude of everyday actions performed both at work and at leisure, that is, “things people will tell you they do if you ask them and they are not applied linguists” (p. 89). Among examples of these everyday actions, Long listed painting a fence, sorting letters, weighing a patient, helping someone across the street, and so forth. Most other SLA researchers do not extend the definition of a task to include language-free activities considering that the overall goal of tasks is to promote language use and development (R. Ellis, 2003). Therefore, only tasks involving the TL were considered in the present study.

Regardless of what exactly is emphasized in each particular definition, a major unifying factor is the presence of communicative language use for a predetermined goal that resembles a real-life function (R. Ellis, 2003; Leaver & Willis, 2004; Nunan, 1989, 1991, 1999; Willis & Willis, 2007). In language pedagogy, tasks are used in order to provide learners with the kinds of experiences they need for the development of true ability to function in the language, rather than for acquiring systematic knowledge about the language. In SLA research, tasks serve as a way of eliciting learner TL speech samples for the purposes of studying the processes involved in language acquisition.

In their meta-analysis of effectiveness of task-based interaction in acquisition of specific lexical and grammatical items, Keck et al. (2006) used Pica, Kanagy, and Falodun’s (1993) simple definition of tasks as activities that engage a pair (or a small

group) of learners in work toward a particular goal. This definition is open to interpretations, especially in light of the fact that misconceptions about tasks abound (Cobb & Lovick, 2007). In this study, tasks were conceptualized primarily by using R. Ellis' (2003) extended definition. This definition was applied in conjunction with the criterial features of tasks presented in the subsequent section to determine whether a treatment used in a primary study selected for the meta-analysis indeed represents a task.

According to R. Ellis, a task can be defined as follows:

a workplan that requires learners to process language pragmatically in order to achieve an outcome that can be evaluated in terms of whether the correct or appropriate propositional content has been conveyed. To this end, it requires the learners to give primary attention to meaning and to make use of their own linguistic resources, although the design of the task may predispose them to choose particular forms. A task is intended to result in language use that bears a resemblance, direct or indirect, to the way language is used in the real world. Like other language activities, a task can engage productive or receptive, and oral or written skills, and also various cognitive processes (p. 16).

Because R. Ellis' (2003) definition is multifaceted, it sometimes is challenging to apply in practice to operationalize the construct of task. So-called criterial features of tasks that were used along with R. Ellis' definition are outlined in the next section.

### *Criterial Features of Tasks*

Even though there is no complete consensus in the SLA field about the concept of task and various authors may emphasize particular aspects of tasks over other aspects, there is a certain degree of agreement about the so-called criterial features that distinguish tasks from nontask activities (R. Ellis, 2003; Willis, 2004). Predominantly, a task is perceived as a piece of work (Nunan, 1993) completed by learners for a genuine, meaningful purpose, rather than for "language display" (i.e., demonstrating that one can express adequately a prescribed utterance in the TL) and has a clear, observable work



product (R. Ellis, 2003; Nunan, 2004, 2006; Willis, 2004). For example, learners can be asked to come up with a joint plan of action, compile a ranked list of arguments, make a prediction, reach consensus on how to resolve a moral dilemma, find discrepancies between two sources of information, and so forth in the TL. The observable product of such activities can be a plan, a list, a chart, a consensus (presented verbally or in writing), and many other outcomes that are found in real-life situations outside the language classroom.

Based on an extensive review of literature, R. Ellis (2003) identified the following six criterial features of tasks (pp. 9-10).

1. A task has a specific workplan (R. Ellis, 2003). This criterion clearly serves to distinguish tasks from free-flowing conversational exchanges. It is supported by, among others, Lee (2000) who emphasized the role of task as a mechanism for structuring and sequencing learners' interaction and Breen (1989) who referred to task as a "structured plan." According to Nunan (1993), a task "should also have a sense of completeness, being able to stand alone as a communicative act in its own right" (p. 59). Breen (1987) explained that task workplans can range from simple to more complex ones that involve group problem-solving or simulations. The duration of task performance can vary from several minutes to a couple of hours of class time based on the learner level, task type, and pedagogical purpose. Frequently, tasks are chained together with each subsequent task building on the outcome of the preceding one (Nunan, 1999), in which case they probably should be viewed as task sequences rather than individual tasks. Lengthier learner activities, especially those that span several class sessions are more likely to fit the definition of a project than a task.

2. A task is an activity where the primary focus is on conveying meaning (R. Ellis, 2003; Nunan, 1989) versus the display of ability to use correct forms to express meaning that has been dictated by someone else such as the teacher or the textbook writer (i.e., “language display”). For this reason, tasks typically incorporate a so-called *gap* such as an *information*, *reasoning*, or *opinion gap* (R. Ellis, 2003; Leaver & Willis, 2004) that are defined in this chapter in the section titled *Task Types*. Tasks cannot require learners to regurgitate other people’s meaning exclusively (Skehan, 1998), and activities that only involve manipulation of language form, rather than meaning, are defined as exercises (e.g., when the learners are asked to change the singular forms to plurals or a story told in the past tense to future tense). Arguably, exercises that manipulate form do not ignore meaning, however, in Widdowson’s (1998) terms, exercises focus on so-called semantic meaning, whereas tasks focus on pragmatic meaning because they require solving a specific communication problem and are assessed in terms of their communicative outcome (Skehan, 1998). In their definition of tasks, Richards, Platt, and Weber (1985) stressed that tasks provide purposes to classroom activities which “go beyond the practice of language for its own sake” (p. 289).

3. A task involves the same processes that are found during language use in the real world (R. Ellis, 2003). R. Ellis distinguished between so-called real-world tasks (e.g., completing a form in the TL) and pedagogic tasks (e.g., finding the differences between two pictures by talking about them with a partner). Whereas the latter activity hardly occurs in the real world in the same format, it arguably gives rise to the same kinds of interaction as real-world tasks (Nunan, 1989).

4. A task can involve any of the four language modalities, that is, skills (R. Ellis,

2003), both receptive (i.e., listening and reading) and productive (i.e., speaking and writing). According to Willis (2004), tasks may entail any number of language skills, from only one to all four, as well as any combination of these skills. Some researchers stress the importance of tasks that involve learner interaction (R. Ellis, 2003; Leaver & Willis, 2004; Lee, 2000; Long, 1985, 1989). Hypothetically speaking, however, a task does not have to include learners' interaction with each other. Richards et al. (1985) gave examples of tasks that do not include language production at all, even by individual learners, for example, drawing the map of an area while listening to someone describe this area or listening to instructions and performing the required actions. R. Ellis (2003) reported that research has shown that, when interaction is required for task completion, negotiation of meaning opportunities leading to language acquisition are enhanced. The present study only deals with tasks that require interaction.

5. A task involves learners' cognitive processes (R. Ellis, 2003) that are used in real life outside language classrooms such as rank ordering, reasoning, evaluating information, and so forth. According to Leaver and Kaplan (2004), to the extent possible, tasks should incorporate cognitive skills that are classified as higher order skills in Bloom's taxonomy (i.e., analysis, synthesis, and evaluation) rather than only lower-order cognitive skills such as comprehension or repetition (Anderson & Krathwohl, 2001) that are present in language exercises.

6. A task has an observable product or outcome (R. Ellis, 2003) that is not the same as the displayed language use. Prabhu (1987) pointed out that a task requires learners "to arrive at an outcome from given information through some process of thought" (p. 2). This real-world product, or outcome, can be a family tree, a plan, an

itinerary, a chart, an advertisement, description of an imaginary product, a letter, a set of instructions created by learners, and so forth. According to Willis (2004), observable outcomes of tasks can be tangible (e.g., a schedule) or intellectual (e.g., solution to a problem). They can be verbal, that is, written or oral, as well as nonverbal such as a drawing, a floor plan, a map, an identified person, and so forth. The presence of a concrete, observable outcome distinguishes a task from free, unstructured-conversation practice in the TL that has a process but not a product.

Adapting R. Ellis' (2003) criterial features, Cobb and Lovick (2007, pp. 8-9) listed the following characteristics that they believe to be most helpful for classroom teachers in determining whether a classroom activity is a task: (a) presence of a workplan, (b) interaction between learners, (c) nonlinguistic purpose for the learners' interaction, (d) manipulation of information and not merely of language form, (e) involvement of cognitive processes that humans generally use in life outside of language learning, (f) connection to real-world events and functions, (g) presence of a predetermined observable product, not merely of the process of interaction, and (h) possibility of multiple outcomes (with the exception of tasks that resemble logic puzzles or mathematical problems such as figuring out the most cost-effective way of completing a project).

Classroom activities also may be viewed on a continuum, with “tasks” and “nontasks” on the opposing ends of the continuum line, and various activities hypothetically may fall on different points of this continuum. Certain communicative classroom activities may meet some or most, but not all, of the requirements for tasks (R. Ellis, 2003). Moreover, some authors extend the concept of task to include so-called

consciousness-raising activities and other types of activities where language itself, rather than real-life information, is the content of the task (R. Ellis, 2003; Fotos, 1994; Pica, 2009). In such tasks, learners talk about specific language features or their appropriate use, and the observable outcome might be a hypothesized rule or a classification of language items that the learners have created based on their discussion with each other. Such tasks are discussed in more detail in the section of this chapter titled *Focused and Nonfocused Tasks*. These types of tasks do not meet the criterion for a nonlinguistic purpose for learners' interaction similar to what happens in the real world outside the language classroom. For this reason, consciousness-raising and other types of metalinguistic activities that aim to raise learners' awareness of linguistic features through discussions *about* the language were not considered to be TL communication tasks in the present study.

In view of these considerations, operationalization of a task in research is quite challenging. For the purposes of distinguishing tasks from nontasks in the present study, the following main criteria were applied: (a) presence of a workplan, (b) presence of nonlinguistic purpose for the learners' interaction, (c) manipulation of real-life information, and (d) presence of a clearly defined real-world observable product (i.e., not merely evidence of TL input comprehension or TL production by the learner).

### *Benefits and Limitations of TBLT*

Some of the benefits of TBLT to language learners were discussed in previous sections. This section briefly summarizes these points and links TBLT with deeper levels of processing that are hypothesized to contribute to long-term retention, as well as with the social aspects of learning. Some known caveats and limitations of TBLT are

presented as well.

TBLT provides a holistic, natural approach to learning the language and helps overcome the inert knowledge problem, that is, learners' inability to make use of their TL knowledge in a real communicative setting in real time (Larsen-Freeman, 2001). TBLT takes into account learners' internal processability constraints (Pienemann, 1984, 1989), nonlinear nature of interlanguage development (Kellerman, 1985; Nunan, 1999; Selinker, 1972), and the natural order of acquisition of language structures that research has shown to prevail over any prescribed textbook order (R. Ellis, 1994a; Kwon, 2005; Pienemann, 1989; Schumann, 1979). As presented earlier, from the skill-acquisition perspective, communication-task completion contributes to the development of automaticity in the use of the TL defined by DeKeyser (2001, 2007) as ability to perform complex tasks quickly and efficiently, without having to give primary focus to many of the linguistic procedures involved (De Ridder, Vangehuchten, & Gomez, 2007). TBLT constitutes "transfer-appropriate processing and other positive features of communicative practices" (Segalowitz, 2003, p. 402). Additionally, TBLT is an intrinsically motivating instructional technique that allows for learners' self-expression in creating the required product and gives the learners a sense of accomplishment (Nunan, 1989; Willis & Willis, 2007).

Gass and Varonis (1989), among other researchers, found evidence of a greater number of negotiation repairs that, according to Long (1996), are conducive to language acquisition during NNS-NNS discourse in tasks as compared with free-conversational practice. A likely explanation is that, in the picture task that was used in Gass and Varonis' study, the learners were pushed to produce utterances conveying more detailed

information to their interlocutors, whereas in free conversation they had a greater degree of control about what messages they attempt or do not attempt to convey.

Robinson (2007) argued that task performance requires learners to engage in complex thought (e.g., ability to reason) as well as to act on their thoughts and adapt to the interactional demands of the task and to the other participants involved in the completion of the task. Therefore, tasks encourage a greater investment of mental effort and create the intensity of use necessary for deeper processing that leads to better encoding of the language material and higher probability of successful subsequent retrieval ( Craik, 2002; Craik & Tulving, 1975). Exercises, as opposed to tasks, barely scratch the surface of the learners' consciousness and contribute more to learning *about* the language (i.e., declarative knowledge) than to true acquisition. Tasks completed in small groups inevitably bring in the affective and social aspects of learning. When learners work together on task completion, they may have to work through the initial confusion and to mobilize their resources to overcome cognitive clashes between themselves and their partners (Vygotsky, 1986). Therefore, the learners tend to internalize the language items used in the process better, which leads to stronger long-term retention. Compared with tasks, such types of activities as language drills, controlled linguistic practice, and even teacher-controlled metalinguistic (i.e., rule-discovery) activities constitute shallow processing and thus can lead only to short-term learning rather than enduring acquisition (Tomlinson, 2007).

In summary, TBLT is a solid, learner-centered, instructional approach that potentially is better aligned with learners' internal syllabi and creates opportunities for deep processing of the target language features and for developing automaticity of their

use. Nevertheless, the research findings regarding the effectiveness of tasks are not necessarily definitive and conclusive. Swan (2005), who admitted that TBLT may help improve learners' command of already known language items, at the same time was rather critical of the notion that task-based instruction is appropriate for the systematic teaching of new language items.

Regarding the claim that learners' interaction in tasks creates greater opportunities for negotiation of meaning and form than free conversation, Nakahama, Tyler, and van Lier's (2001) empirical study provided evidence to the contrary. Long (1996), a strong proponent of TBLT, labeled free, unstructured conversation "notoriously poor" in TL development as compared with tasks. Nakahama et al., however, demonstrated that free-conversational exchanges also create opportunities for negotiation of meaning while at the same time providing greater challenges in maintaining the conversational flow on the discourse level than structured information-gap tasks. Moreover, Foster (1998) found that, in her study setting, the learners did not employ negotiation for meaning strategies during task-based group activities when they encountered gaps in understanding. She drew these conclusions on the basis of observing 21 intermediate-level part-time students of English at a large municipal college in Great Britain complete four classroom tasks. The participants, most of whom were female, came from a wide variety of L1 backgrounds (e.g., Arabic, French, Korean, Spanish) and ranged in age from 17 to 41, with an average age of 24 years.

It is conceivable that in Nakahama et al.'s (2001) study the three high-intermediate ESL learners of Japanese origin, who were college-educated, had studied English for 6 years, were between 25 and 30 years old, and had relatively high scores



(545, 535, 550) on the Test of English as a Foreign Language (TOEFL), were mature and highly motivated. Therefore, they frequently engaged in TL hypothesis testing during free conversation, and when miscommunication occurred, they worked to repair the conversation through negotiation of meaning that learners with different characteristics may not have done in the same way. In general, the effectiveness of communication tasks has been hypothesized to be moderated by a wide range of factors such as the learners' proficiency levels, personal goals, personality factors, familiarity with TBLT, attitudes toward TBLT, presence of pretask planning time, quality of task design, and so forth as presented in the subsequent sections of this chapter. These learner-related, task-related, and context-related variables were coded in the present meta-analysis when the information was available in the included primary studies for the purposes of examining them as potential moderator variables.

The disparate research findings also point to the need for a balanced approach where TBLT is not adhered to in the strictest sense but is used in combination with other classroom techniques and activities. Results of SLA research suggest that, although the analytic approach is better aligned with learners' natural acquisition processes than the synthetic syllabus, it needs to be augmented with more focused grammar instruction (R. Ellis, 2003; Lightbown, 2007; Pica, 2009; Skehan, 1998, 2001). Learners frequently can rely on their strategic communication skills to complete the task without giving proper attention to language form, that is, without trying to make their utterances grammatically appropriate and without fully understanding how the form functions to convey meaning (R. Ellis, 2003; Lightbown, 2007; Pica, 2009). Therefore, some researchers (Skehan, 2001; Skehan & Foster, 2001) caution that exclusive reliance on task-based interaction,

devoid of focused attention to language structure, may result in inadequate grammatical accuracy and fossilization of grammatical errors.

In view of these considerations and challenges associated with TBLT that are discussed in this section, more solid empirical evidence is needed that TBLT can be used effectively in improving grammatical accuracy in the TL. Effective integration of the teaching of grammar with TBLT broadly is the focus of the present study that aims to investigate the effectiveness of learner interaction in specially designed communication tasks that promote the use of specific target grammatical items as compared with other types of FFI (Research Questions 1 and 2). The issue of integration of teaching grammar into meaningful tasks is presented in more detail in sections titled *Focused and Nonfocused Tasks* and *Pedagogical Grammar and Language Acquisition* in this chapter.

#### Types of Tasks as Moderator Variables

Research Question 3 in the present meta-analytic study has to do with the types of communicative tasks that are most effective in facilitating acquisition of FL and L2 grammatical features. Numerous task typologies have been created that examine tasks used in classroom teaching and research from many different perspectives (Willis, 2004). Interactive tasks differ in terms of scope, intended learner levels, pedagogical purposes, cognitive processes involved, characteristics of the interactional flow, degree of difficulty and complexity, whether they target the development of general facility with the language or specific language features, and so forth (R. Ellis, 2003; Robinson, 1998; Skehan, 1998, 2001; Willis, 2004). Because it was not possible to review all existing task classifications, only those task types that potentially could serve as moderator variables in the present meta-analysis were reviewed in subsequent sections of this chapter. Some

other important task types, for which accumulation of sufficient numbers of primary studies was unlikely, are mentioned but not reviewed in great detail.

### *The Gap Principle and Major Task Designs*

One of the most common principles of task design is the so-called *gap principle* (Willis, 2004). Prabhu (1987) identified three types of gap tasks.

1. *Information-gap* tasks are activities where one or more participants hold information that must be given to others in order for the pair or group to be able to complete the task (Prabhu, 1987). A typical example of an information-gap task is an activity where two learners have to figure out the differences between artifact A and artifact B by talking about them but not showing them to each other. The artifacts in this case can be real objects, pictures, maps, schedules, video clips, and so on. So-called *memory-gap* tasks also are based on the gap principle. They rely on the fact that, given a limited amount of time to remember the contents of a picture or a video clip, participants will remember different pieces of information, and, therefore, a natural gap will be created (Willis, 2004). The participants have to exchange information in order to be able to compile a complete description of the picture or a narrative based on the video clip. Some researchers reserve the label information-gap tasks only to describe such activities where information flows from one of the participants to the other (i.e., one-way tasks) but not in both directions (i.e., two-way tasks; Doughty & Pica, 1986; Pica et al., 1993). These researchers refer to two-way tasks (where information is held by two or more participants and all of the information is required to complete the task) as *jigsaw* tasks.

2. *Reasoning-gap* tasks are activities where all participants have access to the same information upfront but must reason collaboratively in order to deduce an outcome

from the given information (Prabhu, 1987). The key here is that a solution or a decision has to be reached through interaction among the participants. Reasoning-gap tasks typically fall into two categories: (a) *problem-solving* reasoning-gap tasks (e.g., students collaboratively figure out a person's entire weekly schedule of activities based on the provided class schedules, employee schedules at the place of the person's employment, the person's desk calendar, etc.) and (b) *decision-making* reasoning-gap tasks that call for a pair or a small group to agree on a course of action in a given situation (Willis, 2004).

It is often pointed out that in reasoning-gap tasks participants do not have to interact to complete the task successfully (Keck et al., 2006), for example, learners may choose to work individually to solve the problem presented by the task. Therefore, it may require special effort on the part of the teacher to encourage learner-to-learner interaction if such interaction is desired in a reasoning task. Doughty and Pica (1986) reported that tasks in which the information exchange between the participants was required generated a considerably greater amount of modified interaction than tasks in which the exchange of information was optional, and this difference was statistically significant.

3. *Opinion-gap* tasks are activities that require participants to formulate their opinions, typically on a societal issue (Prabhu, 1987). Because opinion exchanges frequently occur in the format of free-flowing conversation that does not meet the criteria for a task, it is important that such activities are designed to have more structure (i.e., a workplan) and a clear observable product. For example, opinion-gap tasks can require participants to justify their point of view, evaluate the ideas of others, find commonalities and differences between their own position and that of their partners, rank brainstormed ideas, and so forth (Willis, 2004; Willis & Willis, 2007).

In addition to the three types of gap tasks (i.e., information-gap, reasoning-gap, and opinion-gap tasks), some researchers differentiated so-called *information-transfer* tasks (Willis, 2004) such as those requiring reproduction of written or spoken information in some graphic form (e.g., a chart or a table) and vice versa (i.e., information presented in a table, chart, or drawing reproduced in the form of a narrative). Another example of information transfer is a task in which participants are required to listen to a narrative and present the information they heard in the form of an interview, a radio commercial, a brochure, a bulleted list of instructions on a 3 by 5 card, and so forth. For example, Revesz (2007) had participants create a narrative of events based on a photo, and Revesz and Han (2006) based on a video or notes.

Keck et al. (2006) also defined *narrative* tasks as a special category if, while the speaker tells a story, the listeners are required to interact by providing feedback in the form of recasts and requests for clarification. In the present study, personal narratives that are not outcomes of an information-transfer task were included only if they met the required criteria for tasks as outlined in the section titled *Criterial Features of Tasks*, for example, if an observable product is present. In this case, such a narrative task can be considered an information-gap task where one or more of the partners have to share information unknown to the other(s). Without an observable outcome (e.g., the listeners drawing a picture representing what they heard or preparing a list of similarities and differences between what they heard and their own experiences), such personal or other narratives (e.g., retelling of a story), including those where the listeners were encouraged to interact with the narrators, were considered free conversation because they have a process but do not result in a task product as defined by R. Ellis (2003).

Some researchers specified *simulations* or *role-play* tasks such as, for example, booking a flight or returning a piece of purchased clothing as a separate task category (Willis, 2004). In such tasks, the make-believe communicative outcome (i.e., the flight has been booked or the clothing item has been returned) satisfies the requirement for the presence of the observable product.

The gap principle is a useful principle of task design; however, many classroom tasks are compound in nature and cannot be assigned easily to one of the gap categories (R. Ellis, 2003). For example, a task can start out as a jigsaw task with the necessary pieces of information split between the partners; however, once the information is exchanged, the partners have to reach a conclusion or a decision based on all available data. For example, one partner may be provided with a schedule of festivities in city X, whereas the other one is given access to the city shuttle schedule. Together, the learners have to agree on the events they would most like to attend and figure out the most efficient way to get around town to see as many of these events as possible. An example of a jigsaw task combined with an opinion-gap task is the activity where the participants are each given a picture of a food pyramid (e.g., for the US and the target culture). After they exchange information with each other, they have to formulate and present a joined position as to the respective health benefits and disadvantages of the eating habits in the two cultures.

Keck et al. (2006) reported that jigsaw and information-gap tasks are by far the most frequently-used classroom tasks targeting specific lexical and grammatical TL features. These types of tasks comprised 90% of all treatment-group tasks in their meta-analysis. A likely reason for this trend is that these types of tasks are considered to be

superior to other types because they push the students to find ways to convey precise information to their partners (Pica et al., 1993). Keck et al.'s (2006) meta-analytic findings support this assertion. Both jigsaw ( $d = .78$ ) and information-gap ( $d = .91$ ) tasks were found to be more effective than the control and comparison conditions in their meta-analysis. The meta-analysts stressed, however, that greater accumulation of studies utilizing individual task types is required before more reliable interpretations of the results can be made.

The following task-type classification was used in the present study: (a) information-gap (one-way), (b) jigsaw, (c) problem-solving, (d) decision-making, (e) opinion-gap, (f) information-transfer, (g) role plays, (h) narratives, and (i) compound tasks (e.g., information-gap and decision-making) when it was not possible to assign a task to only one category. Keck et al. (2006) who used a somewhat different classification in their meta-analysis reported including primary studies with the following task types used as treatments: (a) jigsaw, (b) information-gap, (c) problem-solving, (d) opinion exchange, and (e) narrative. Even though Keck et al.'s classification also included decision-making tasks, none of the studies in their meta-analysis used this type of task. As anticipated, due to their underutilization in the field, some of the task types outlined for the present study were not represented even minimally in the candidate primary studies (see *Research Synthesis* in chapter IV).

#### *One-Way and Two-Way Tasks*

Even though the distinction between one-way and two-way tasks was addressed briefly in the previous subsection, it needs to be presented in more detail because of the presence of research studies that specifically investigate this dimension of tasks. The

distinction between one-way and two-way tasks has to do with how the information is expected to flow in a pair or a small group of learners engaged in completing the task. An example of a one-way task is an activity in which one student in each pair is given a map with two locations marked: (a) the current location and (b) the destination. This student is required to give directions to the destination point to the partner who is holding an unmarked map. The partner follows the directions to the destination point but is not required to provide any information in this task. The distinction between the one-way and two-way tasks typically is applied to information-gap tasks; however, it is feasible for this distinction to be made in some other types of tasks as well.

Gass and Varonis (1985) used a picture-drawing task and an information-gap detective story to compare the negotiation of meaning episodes that one-way and two-way tasks generate. The one-way group completed the picture-drawing task by having the designated speaker in each group provide instructions to the listener as to what to draw. In the two-way detective-story task, each of the two participants had information about the committed crime that the other one lacked, so the participants had to exchange information. Gass and Varonis found that, in the one-way picture-drawing task, there were more instances of original input being unaccepted by the listener (therefore requiring elaboration) compared with the number of instances of input being unaccepted by both interlocutors together in the two-way detective story, even though the difference was not statistically significant. The researchers, therefore, concluded that because there is a greater shared background in two-way tasks than in one-way tasks, there are fewer opportunities for communication breakdowns, and, consequently, less need for negotiation of meaning to occur.



Long (1981), however, presented an entirely different finding that two-way tasks lead to more negotiation of meaning and are, therefore, more useful than one-way tasks. In his investigation of the characteristics of NS-NS and NS-NNS interaction, Long randomly assigned 48 adult NSs and 16 adult NNSs of English to 32 dyads (16 NS-NS and 16 NS-NNS dyads). He found that two-way tasks in which all participants had unique information to contribute stimulated a greater number of modified interactions than one-way tasks and that this difference was statistically significant.

It appears that the distinction between the one-way and two-way tasks is not straightforward (R. Ellis, 2003), and many additional factors possibly come into play in addition to (and possibly interacting with) the task type. For example, some researchers asserted that tasks that are designed to create too many communication breakdowns for learners to repair can be discouraging and demotivating as well as lead to error-laden, low-quality interaction (Aston, 1986). Although interaction that occurs in one-way versus two-way tasks has been investigated in a number of descriptive studies, few researchers have attempted to investigate the actual learning in which this interaction results (R. Ellis, 2003). More empirical data clearly are needed. In the present meta-analysis, the distinction between one-way and two-way tasks was investigated as one the moderator variables potentially influencing the learning of specific target structures.

### *Closed and Open Tasks*

Another possible moderator variable that is related to task design is closed versus open tasks. Long (1989) hypothesized that so-called *closed* tasks that have one predetermined correct solution are more beneficial to TL development than *open* tasks for which a wide range of solutions, unique to each pair or group of learners, can be

accepted. An example of an open task is a list of possible pro and con arguments put together by a group of students on a controversial proposal. Closed tasks are designed so as to force learners to work out a single possible solution. For example, the single correct answer for a spot-the-differences picture task is the list that correctly identifies all the differences created by the illustrator. An example of a closed reasoning-gap task is an activity in which the students are required to figure out the seating arrangement for invited party guests where the provided list of the guests' seating preferences that have to be met allows for only one possible configuration.

Long (1996) cautioned that open tasks may resemble free conversation where participants can address topics very briefly and drop them as soon as problems in communication arise. In this case, valuable opportunities for negotiation of meaning and form will be missed and the learners will not be pushed to stretch their current language abilities. This concern was supported by Loschky and Bley-Vroman (1993) who believed that closed tasks promote more negotiation of meaning, focus learners' attention on form, and are therefore preferable to open tasks. In addition to the finding that different task types stimulate different interactional patterns, Nunan (1991) reported that some task types may be more appropriate than others for learners at particular levels of proficiency. For example, closed tasks may stimulate more modified interaction than open tasks at higher levels, whereas for lower levels they may be too challenging and frustrating. Keck et al. (2006) did not report findings pertaining to the effectiveness of closed versus open tasks in their meta-analysis. In the present meta-analysis, this task feature (i.e., closed vs. open) was coded and its possible moderating effects were investigated.

### *Divergent and Convergent Tasks*

In addition to the distinction between closed and open tasks that is made on the basis of the intended task outcome, researchers sometimes differentiate between so-called *divergent* and *convergent* tasks based on the goal-orientation of task participants (R. Ellis, 2003). This classification was proposed by Duff (1986) who defined convergent tasks as those in which learners share the goal of jointly finding a solution that is acceptable to all participants such as in most problem-solving tasks. A traditional example is the well-known “desert-island” task in which learners must agree on a limited number of objects that they can take with them as a group to survive on an uninhabited island.

As compared with the “desert-island” consensus-reaching task, divergent tasks present learners with independent or even opposite, rather than common, goals to accomplish (Duff, 1986; R. Ellis, 2003). A divergent task may include presenting arguments where the participants are assigned differing viewpoints on an issue and have to defend their position and refute their counterparts’ arguments.

Duff (1986) reported that convergent problem-solving tasks appear to be more effective than divergent tasks based on the nature of interaction that occurs between the participants; however, divergent tasks result in greater amounts of learner-produced TL output. Skehan and Foster (2001) pointed out that, in general, whenever there is optionality in the information exchange by the learners, the number of negotiation of meaning episodes is reduced and that more negotiation will occur when the interactants have convergent, rather than divergent, goals and one acceptable outcome, rather than many. Keck et al. (2006) did not report meta-analytic findings relevant to divergent

versus convergent tasks. In the present meta-analytic study, this task feature was coded and investigated as a possible moderator variable.

### *Focused and Nonfocused Tasks*

The present meta-analysis investigated the effectiveness of so-called *focused* communication tasks. A task can be designed so that attaining the prescribed goal depends upon a high degree of linguistic precision that, in turn, requires the use of specific language features referred to as target language features (R. Ellis, 2003; Fotos, 2002; Larsen-Freeman, 2001a, 2003; Nassaji, 1999). These target features can be morphological, syntactic, lexical, or even pragmatic in nature (R. Ellis, 2003). Tasks that are designed to elicit the use of the target items while the students interact for a real communicative purpose frequently are referred to as *focused* tasks as opposed to *nonfocused* tasks that serve to develop general communicative ability in the TL but do not necessarily predispose students to the use of specific TL items (Fotos, 2002; Nassaji & Fotos, 2004).

Some researchers, most notably Skehan (1998), questioned the validity of the notion that communicative tasks can be created so that they require use of prescribed language items. In Skehan's view, by their very nature, tasks are activities in which students are free to use any language resources that they have acquired so far and that predisposing learners to using specific TL items undermines the naturalness of a task. Other researchers (R. Ellis, 2003; Fotos, 2002; Nassaji, 1999) believed that targeted focused tasks are legitimate curricular units and are very beneficial to interlanguage development and prevention of fossilization (i.e., persistent presence of incorrect, nontargetlike forms in the learner's interlanguage; Long, 2006; Selinker, 1972).

Lightbown (2007) commented that it is acceptable to target specific TL items in tasks and to make the linguistic objectives of the task known to learners upfront as long as the teacher moves quickly from discussing the target structure to creating a specific context, or “semantic space,” for its meaningful use.

In case the target of a focused task is grammatical in nature, such a task is sometimes referred to as a grammar-based communication task or an implicit structure-based interactive task (Fotos, 2002). Larsen-Freeman (2001a, 2003) even coined a special term “grammaring” to refer to truly communicative, task-based practice of specific target grammatical items. Designing such focused tasks that promote genuine interaction often requires a great degree of skill and creativity and is often challenging for classroom teachers who find it easier to create tasks that are more global in nature (Larsen-Freeman, 2001a; Cobb & Lovick, 2007).

Larsen-Freeman (2001a) provided the following examples of ESL grammaring tasks: (a) students correct inaccurate factual statements in a story told by the teacher about a well-known recent sporting event while practicing English negation and (b) students “place bets,” or make predictions, about whether a tower built from blocks will fall if additional blocks are placed on top while using English modal verbs and phrases expressing supposition or probability (e.g., “might fall,” “is likely to fall,” “will most definitely fall,” etc.). Other common examples of structure-based communication tasks are information-gap “spot-the-differences” picture tasks where the differences between the pictures are designed so as to predispose students to the use of specific language items (e.g., prepositions and adverbs of location).

Even though “spot-the-differences” tasks appear to be used most commonly in TBLT-related research, focused tasks of many different types can be designed for classroom use and research purposes. For example, to practice expressions of frequency (e.g., “never,” “once a week,” “every Saturday,” “two times a month,” etc.) that are very complicated grammatically in some languages such as Russian, students can create, administer, and report findings of a survey about how often their classmates complete various household chores. To practice conditional mood forms (e.g., “if we were” or “if we could”), students can be asked to reach and report their consensus about various hypothetical situations.

Some researchers proposed the use of tasks where linguistic features themselves become the content of the task as learners try to hypothesize about the grammatical rule, figure out how the structure functions under different conditions using their L1 or TL, and so forth (R. Ellis, 2003; Fotos & Ellis, 1991; Pica, 2009). Although such problem-solving, consciousness-raising, and other types of metalinguistic tasks are undoubtedly useful in language teaching, they do not represent the type of communication tasks in which learners manipulate and convey nonlinguistic real-world information through the use of the TL. For this reason, this subtype of focused tasks was not considered in this study.

In their meta-analysis, Keck et al. (2006) specifically investigated the effectiveness of task-based interaction in focused tasks targeting the development of predetermined lexical or grammatical features. They reported that both grammar-focused ( $d = .94$ ) and lexis-focused ( $d = .90$ ) tasks produced large main effects; however, the 95% confidence intervals were wide for lexis (0.40-1.40). In the present meta-analysis, only

the effectiveness of focused tasks that target grammatical language items, that is, various grammatical forms and structures, both morphological and syntactic in nature, were investigated. Table 2 summarizes the characteristics of tasks that have been presented in the various subsections of the *Types of Tasks as Moderator Variables* section.

Table 2

## Summary of Task Characteristics

Major Task Design (The “Gap” Principle)	Flow of Information	Open-endedness	Participants’ Goals	Presence of Linguistic Target
1. Information gap (information-gap and jigsaw tasks)	1. One-way tasks	1. Closed tasks (i.e. one possible outcome)	1. Divergent (i.e., differing goals)	1. Focused tasks
2. Reasoning gap (problem-solving and decision-making tasks)	2. Two-way tasks	2. Open-ended tasks (i.e., more than one possible outcome)	2. Convergent (i.e., the same goals)	2. Nonfocused tasks*
3. Opinion gap				
4. Other (information- transfer, narrative, role-play, compound tasks)				

\* Nonfocused tasks were not considered in the present meta-analysis.

The task type probably is the most influential factor affecting the nature of interactional negotiation in tasks; however, learner-, teacher-, and context-related variables also are believed to play a role. The next section presents an overview of learner-related variables.

## Role of Individual Learner Differences in Task Performance

Such factors as age, gender, L1, proficiency level, and so forth also affect learners’ performance in tasks and, consequently, how much TL development results from task-based interaction. For example, Gass and Varonis (1985) pointed out that gender plays a role because men tend to indicate lack of understanding during interaction

more frequently than women. Pica, Holliday, Lewis, Berducci, and Newman (1991) reported a similar result, although in their study there was also a cultural factor present. The researchers found that, in their study, the female Japanese student was more passive during the interaction in the sense that she rarely asked her male counterpart for clarification.

In terms of the learner proficiency levels, more proficient learners have been reported to be both more willing and better equipped to address grammar issues during task performance (Williams, 1998). Porter (1986) suggested that the manner in which intermediate and advanced learners, rather than beginning learners, negotiate meaning may be close to the negotiation of meaning that occurs among native speakers of the TL. Seol (2007) who investigated the effect of the degree of linguistic similarity between the learners' L1 and L2 reported that, based on the participants' scores on grammaticality judgment posttests, acquisition in prepubescent learners did not depend on L1-L2 similarity, whereas in postpubescent learners, acquisition largely depended on L1-L2 similarity.

The educational setting, institutional expectations, and characteristics of a program may play a role and interact with learner characteristics as well. For example, Mackey and Goo (2007) observed that the effects for interaction were greater in FL contexts than in L2 contexts. Spada and Lightbown (2008b) pointed out that this finding may be explained by the fact that FL students have fewer opportunities to use the language outside of class than L2 students, and, therefore, FL teaching generally tends to be more form-focused than L2 teaching.



In terms of the learners' cognitive characteristics, Fujii (2005) examined the relationship between three variables: (a) learners' FL aptitude, (b) learners' so-called orientation to form, and (c) their L2 production. Findings indicated a relationship between specific aptitude components and two dimensions of task performance: (a) accuracy and (b) so-called lexical sophistication. Fujii reported that the qualitative analysis she had performed suggested that learners with a greater capacity of the working memory may be able to perform better in complex, multidimensional language tasks.

A widely accepted simplified model of language aptitude includes so-called linguistic analytic ability (i.e., ability to make inferences and generalize about structural encoding of meaning), phonetic coding ability (i.e., ability to discriminate between and retain TL sounds and pitch variations as well as ability to associate them with meaning), and associative memory (i.e., ability to retain in memory associations between TL verbal stimuli and their real-world referents; Skehan, 1998). Some researchers (Carroll, 1990) split the component referred to as linguistic analytic ability into two subsets: (a) grammatical sensitivity (i.e., ability to recognize grammatical and syntactic functions that words fulfill in a sentence) and (b) inductive language-learning ability (i.e., ability to infer structural patterns from the language input, to induce rules, and to make predictions about how new language material may be encoded on the basis of identified patterns).

A variety of affective and personality-related variables play a role in successful task performance as well. For example, Cameron and Epling (1989) found that, when students characterized as active were paired with each other or with more passive students, such pairs performed better in problem-solving tasks than pairs consisting of two passive students. Swain and Lapkin (1998) pointed out that individual students can

approach the same tasks differently, possibly for a whole host of reasons. Therefore, students with different characteristics benefit from the same tasks differentially.

According to Dornyei (2002), so-called task motivation plays a role. The task motivation is influenced by the learner's own characteristics, features of the learning environment, features of the task, the learner's personal goal-setting for the particular task, and the learner's beliefs about effectiveness of TBLT. Some researchers (Coughlan & Duff, 1994) argued that the actual activity resulting from any given task is necessarily co-constructed by the participants on each occasion, thus rendering any accurate predictions of how the task will be performed virtually impossible.

There are potentially many learner-related factors that influence attitudes toward tasks, success in task performance, and ability to profit in the short and long term from task completion. It was not anticipated that a considerable number of primary studies included in the present meta-analysis will have documented such learner-related differences with sufficient detail; however all relevant information about the participants both in the treatment and the control or comparison groups that was available in the included studies was coded and examined.

#### Other Task-Related Moderator Variables

In addition to task types and learner characteristics presented in previous sections, there is a host of other variables that potentially have an effect on how learners benefit from completing focused communication tasks. These variables can be related to the characteristics of the teacher who is conducting a particular task in class as well as to task origin, task complexity, task difficulty relative to the level of the learners, various conditions of task performance, and so on. Some of the major additional factors that

affect language learners' ability to benefit from task-based interaction are presented in this section.

### *General Considerations*

Just as there are variations in task performance based on many learner-related variables, there are variations in how individual teachers set up and carry out the same tasks in class, which affects to what degree individual learners benefit from these tasks (Samuda, 2007). The teacher-related variables affecting task implementation in class are teaching experience, familiarity with the FoF approach, metalinguistic grammatical knowledge, and so forth (Elder, Erlam, & Philp, 2007). Using classroom transcripts of the same "radio-news-bulletin" task being implemented by three different teachers, Berben, Van den Branden, and Van Gorp (2007) reported striking differences in the degree of control afforded the students, in the degree of attention to form as well as in the overall success of task completion. The researchers documented breakdowns in task performance and, in one instance, evidence of the teacher essentially delivering the task outcome instead of the students. Deconstructing the three lessons and the subsequent interviews with the teachers, the researchers presented a discussion of how teachers reconstruct a given task based on their own personal beliefs about TBLT as well as their own needs, skills, teaching styles, the context in which they operate, and their perceptions of their students.

Other variables that potentially can affect success of the outcome of a classroom task relate to the conditions under which the task is performed. For example, Crookes (1989) and Foster and Skehan (1996), among other researchers, reported that learners tend to produce longer and syntactically more complex output strings in tasks requiring

monologic TL production under the so-called pretask-planning condition, that is, if they are allowed planning time before task interaction begins. The learner's familiarity with the format of a particular task type is believed to play a role as well, however, more research is needed to examine the relative effects of asking learners to perform the same task (i.e., task repetition) versus asking them to perform a similar task (i.e., task familiarity; R. Ellis, 2003). Finally, interlocutor familiarity (i.e., the learner's familiarity with the interaction partner or partners) has been pointed out as another factor affecting task performance (R. Ellis, 2003). R. Ellis explained that being familiar with one's partner, on the one hand, can make it easier for a learner to ask for clarification but, on the other hand, it may reduce the amount of negotiation for meaning if interlocutors are very familiar with each other's voices and TL interlanguage (R. Ellis, 2003). Finally, Samuda (2007) pointed out substantial differences in quality of task design between, for example, tasks designed by classroom teachers and those designed by skilled curriculum developers.

The above-mentioned factors undoubtedly affect the success of task-based interaction in the classroom. Even though significant accumulation for these variables in primary research was unlikely, every effort was made to document and examine the effects of these variables if they were reported in the primary studies included in the present meta-analysis. Where sufficient data needed to calculate associated effect sizes were not available, this information served as a basis for formulating suggestions for future research directions (see the *Recommendations for Research* section in chapter V).

#### *Learner-to-Learner versus Teacher-led Interaction*

A factor of critical importance that has a potential of greatly influencing the

outcome of task-based interaction is whether the interaction takes place among learners themselves or includes an NS participant. Long's (1996) interaction hypothesis posited that conversational interaction can promote TL development in productive ways.

Numerous empirical studies have provided evidence that development of TL morphosyntax and lexis is facilitated by interactions between NSs of the TL and NNS learners (Gass & Varonis, 1994; Mackey, 1999). For many learners, however, the majority of their TL interactions occur in conversation with other learners (i.e., NNSs) rather than with NS interlocutors (i.e., teachers, TAs, or NS tutors). Studies that have contrasted learner-to-learner interactions with NS-NNS interactions have found that these processes differ in substantial ways (Bruton & Samuda, 1980; Gass & Varonis, 1989; Mackey, Oliver, & Leeman, 2003). Because of these differences, it is unclear whether the beneficial effects of NNS-NNS interactions are similar to the beneficial effects of NS-NNS interactions. Wigglesworth (2001) believed that beneficial negotiation of meaning is more likely to take place in NNS-NNS interaction because the more limited nature of the shared background between partners in such dyads leads to greater frequency and complexity of negotiation episodes. Toth (2008), however, reported higher results on the posttests for teacher-led groups than for learner-to-learner interaction groups after task-based grammar teaching. He explained that the interaction transcript data suggested that teacher-led discourse in tasks provided an opportunity for learners to benefit from the teacher's guidance and efforts to direct their attention to the target structures.

Such disparate findings make the variable of learner-to-learner (i.e., learner-led) versus teacher-led interaction during task completion potentially a very important moderator variable in studies of the effects of task-based interaction. As stated in chapter

I under *Limitations of the Previous Meta-Analyses*, Keck et al.'s (2006) meta-analysis included only three studies where NNS learners interacted with each other. One of the reasons that teacher-led interaction produces better results could lie in the characteristics of corrective feedback, or error correction techniques, that the teacher employs, both explicit (e.g., provision of correct form or metalinguistic comments) and implicit (e.g., recasts or requests for clarification; Lyster & Ranta, 1997).

The effects of different types of corrective feedback provided by the teacher were not investigated specifically in this meta-analysis. It is assumed that, in real classroom settings (vs. so-called laboratory settings), whether teachers participate in task completion along with the learners or only monitor learner-to-learner task interaction, they make situated, informed decisions about what, when, and how to correct. The decisions teachers make depend upon the pedagogical purpose of each particular activity, their understanding of the learners' goals, needs, cognitive styles and preferences, and many other factors including learner proficiency levels, group dynamics, teacher's own preferences and strengths, and so forth (Brown, 2001; Byrd, 1998; R. Ellis, 2003, 2007; Kim & Han, 2007; Lightbown & Spada, 1993). As Prabhu (1987) pointed out, in actual classrooms (vs. prescribed lab conditions), error correction during tasks typically is incidental, rather than systematic, in nature. For the purposes of the present meta-analysis, only the absence or presence of learners' error correction during task completion and the absence or presence of error treatment in the posttask phase was recorded if the primary study report provided such information.

#### *Cognitive Complexity of the Task*

SLA literature distinguishes between *task difficulty* and *task complexity*. *Task*

*difficulty* generally refers to the learners' perceptions of the cognitive demands of a particular task that are determined by the learner's proficiency level and aptitude as well as such affective variables as confidence and motivation (Robinson, 2001b). *Task complexity* refers to the "attentional, memory, reasoning, and other information processing demands imposed by the structure of the task on the language learner" (Robinson, 2001a, p. 29).

Robinson (2007) pointed out that early attempts to classify L2 classroom tasks listed such task characteristics potentially affecting learner performance as cognitive load, communicative stress, and so forth. He proposed three broad criteria: (a) interactional criteria (e.g., the number of participants, whether a convergent solution is required, etc.), (b) cognitive criteria (i.e., resource-directing variables such as whether perspective-taking, spatial, causal, or intentional reasoning are required, and resource-dispersing variables such as how many steps the completion of the task entails, whether the participants are given planning time, etc.), and (c) ability-determinant criteria (i.e., high vs. low demands on the working memory, etc.).

Robinson (2007) believed that the learner's attentional resources are not limited at any given time and that it merely is a matter of executive control where the learner chooses to direct his or her attention (vs. being a matter of limited attentional resources). He, therefore, argued that increasing task complexity by raising its cognitive, or conceptual, rather than procedural demands, will lead the learners to "complexify" their language output and, therefore, will lead to development of greater grammatical accuracy.

The issue of task complexity is a controversial one. Robinson's (2007) position has given rise to a number of research studies based on his model; however, numerous other studies have provided evidence that attention during learning is indeed of limited capacity, and, therefore, attending to one source of information inevitably detracts from utilizing another source (Kruschke, 2005). Van Patten (1990) as well as Skehan and Foster (1998, 2001) assumed a limited capacity model of attention that is based on an assertion that humans have limited information-processing capacity at any given time. The implications for L2 learners are that there is a constant competition for their attentional resources between form and meaning and that, therefore, more cognitively demanding tasks will require more attention to the content so the learners will have fewer resources left to attend to TL input and output (Skehan & Foster, 2001). Skehan (1998) cautioned that, due to scarcity of attentional resources that the learners have at their disposal, increasing task complexity most likely will decrease accuracy of production. This consideration points to the need for balancing the requirements for accuracy, fluency, and task complexity carefully in task design.

Skehan and Foster (2001) performed a post hoc meta-analysis of six datasets to investigate whether the accuracy, fluency, and linguistic and structural complexity of learners' TL output are affected by various dimensions of task performance. The tasks in the six primary studies were classified along the following dimensions: (a) familiarity of information involved in the task to the participants, (b) dialogic versus monologic nature of the task, (c) the degree of task structure (i.e., whether a clear, sequential macrostructure for the expected speech event was present), (d) cognitive complexity of outcomes (i.e., whether straightforward decisions vs. multifaceted judgments were



required), and (e) transformation (i.e., whether the participants were required to operate on the information in some way or simply to reproduce it). Skehan and Foster's most important findings were (a) the fluency of the learners' TL output increased with greater task structure, (b) greater linguistic complexity of TL output was present when the task set expectations for a cognitively more complex outcome, and (c) greater linguistic complexity of TL output was present when transformation of information was required under the planned condition (i.e., presence of pretask planning).

Attentional manipulation of tasks in language learning research is a challenging undertaking. Nevertheless, every effort was made in the present meta-analysis to document dimensions of task complexity when they were reported by the primary researchers or could be inferred from the study reports. In general, in line with Research Question 4, all task-related, learner-related, and context-related characteristics that can be obtained or inferred reliably from primary study reports were recorded during the meta-analysis using the researcher-designed coding form (see Appendix C). There were some minor modifications of the coding form, the need for which was identified during the coding process and the discussions with the second coder as presented in chapter III in the *Coding* section.

### Pedagogical Grammar and Language Acquisition

This section provides a situated presentation of task-based interaction in focused (structure-based) TL communication tasks, that is, the independent variable in the present meta-analytic study, as one of the so-called Focus on Form (FoF) instructional techniques. Additionally, a discussion of possible interaction between the type of the specific grammatical structure and the effectiveness of task-based interaction that targets

acquisition of this structure is presented. The concept of *task-essentialness* of the target structure that was treated as a major moderator variable in Keck et al.'s (2006) meta-analysis is introduced as well.

### *Focus on Forms, Focus on Form, and Focus on Meaning*

This subsection provides an in-depth presentation of the Focus on Form (FoF) approach as a methodological principle of TBLT that was introduced briefly in chapter I. A review of its alternatives, Focus on Forms (FoFS) and Focus on Meaning (FoM), is presented in order to delineate the differences between the three approaches clearly.

#### *Focus on Forms (FoFS)*

For thousands of years, studying an FL or L2 primarily consisted of grammatical analysis of TL sentences and translation of TL written forms and texts into L1 (Howatt, 1984; Macaro, 2003; Rutherford, 1987). Originally developed for the analysis of the Greek and Latin languages, this method later was transferred to teaching of English as an FL or L2 and focused on the study of the parts of speech (i.e., nouns, verbs, adjectives, etc.), their morphological variations, and the associated rules. This instructional approach was still dominant in the US and England in the first part of the 20<sup>th</sup> century and is still used in many English Foreign Language (EFL) classrooms throughout the world (Hinkel & Fotos, 2002).

This traditional approach has been referred to by Long (1991, 1996, 1997, 2000) as *Focus on Forms (FoFS)*, that is, analysis of isolated discrete forms outside of real, authentic communication in the TL. It is based on a synthetic syllabus where the teacher or the textbook writer divides the TL into segments of various kinds (e.g., phonemes, words, morphemes, sentence patterns, tones, etc.) and presents these items to the learner

through rules and models, one item at a time, in a sequence determined by the perceived difficulty of these items (Doughty & Long, 2001; Long & Robinson, 1998). Learners are expected to master the correct use of these segments one after another to perfection (Nunan, 1999) and eventually synthesize them for communication. Because it is impossible to locate authentic texts that would contain only the set of items that the learner currently knows, most TL texts used in the FoFS approach are written by the course designers and, therefore, are devoid of many natural features of rich, authentic TL discourse (Parker & Chaudron, 1987; Yano, Long & Ross, 1994). These artificial limitations on what grammatical forms, lexis, and discourse features the learners can experience leave them relying on limited and contrived TL input in the process of learning the TL linguistic code (Cobb, 2004).

Typical classroom activities in the FoFS approach are explicit grammar rule explanations and recitations, repetition of models, memorization of short dialogs, reading of linguistically "simplified" texts, transformation exercises that manipulate form, explicit error correction, and answering of display questions in the TL (i.e., questions to which the inquirer already knows the answers; Long, 1997). There is very little, if any, communicative use of the language.

FoFS as an exclusive instructional approach to teaching TL suffers from a number of serious shortcomings. It is a "one-size-fits-all" approach (Long, 1997) because students' learning styles and preferences are not taken into account, that is, it is assumed that any type of learner should be able to benefit fully from abstract rule explanations and repetitive pattern drills (Cobb & Lovick, 2008). Synthetic syllabi, in general, ignore research findings that acquiring new items is never a one-time event (Nunan, 1999); in

fact, the synthetic approach assumes that SLA is equivalent to accumulating language entities (Rutherford, 1987). FoFS adheres to the obsolete and discredited behaviorist model of SLA (i.e., forming the habit of providing specific TL responses to various TL stimuli in the learners; Long, 1997). Contrary to the assumptions underlying FoFS, research findings have shown that acquisition sequences do not reflect textbook instructional sequences (R. Ellis, 1989; Lightbown, 1983; Schumann, 1979) and that teachability of a TL item at any given point is constrained by learnability (Pienemann 1984, 1989).

Additionally, due to lack of engagement in true communicative tasks, FoFS frequently demotivates students who are not interested highly in linguistic analysis (Long, 1997). This approach leads to the so-called inert knowledge problem (Larsen-Freeman, 2001b), when the learners are not able to apply their knowledge of grammar gained in decontextualised classroom exercises to spontaneous TL use situations.

### *Focus on Meaning*

The opposite extreme is the exclusive *Focus on Meaning (FoM)*, that is, the kind of teaching that occurs in immersion or content-based language-education programs where no teaching of structure is ever deliberately attempted (Doughty & Williams, 1998; Long, 2000). This approach stems from the belief that most of L2 learning is not intentional but incidental and implicit, just like the learning of L1 (Long, 1997). Therefore, the teachers' task is viewed primarily as recreating natural conditions under which the students acquired their L1 in early childhood. The TL is taught through purely communicative activities that are, in this case, based on rich, authentic TL input that makes this option attractive to many learners.

FoM represents an analytic, rather than a synthetic, approach to syllabus design because learners are expected to figure out subconsciously how the language works. Although Long (1997) considered FoM to be a great improvement over FoFS, research findings have provided ample evidence of the limited effectiveness of this approach for developing adequate structural accuracy (Long, 1996, 2000; Lightbown & Spada, 1993). Swain (1991) pointed out that, even after 12 years of classroom immersion in the Canadian French immersion programs, students' productive skills remained far from native-like, particularly with respect to grammatical competence; for example, learners of French failed to mark articles for gender. Moreover, there is increasing evidence of the presence of the so-called maturational constraints for adult language acquisition (Curtiss, 1988; Hyltenstaam & Abrahamsson, 2003; Long, 1990, 1993; Newport, 1990), that is, the fact that postpubescent adolescents and adults regularly fail to achieve native-like TL levels. This lack of ultimate attainment (i.e., native-like proficiency levels) are not due to lack of motivation or ability on the part of adult (i.e., aged 13 and older) learners but rather due to the fact that, around the onset of puberty, humans are believed to lose access to innate language-acquisition abilities they possessed in early childhood and utilized in learning the L1 (Hyltenstaam & Abrahamsson, 2003).

Many researchers agree that a pure FoM approach is insufficient because, contrary to Krashen's (1985) assertion, comprehensible TL input is a necessary but not a sufficient condition for TL acquisition (Gass, 1988; Long, 1997; Swain, 1985, 1991; White, 1987, 1991). Lacking or inadequate attention to the TL grammatical structure leads to premature TL stabilization in adult learners, including fossilization of faulty grammatical structures (Higgs & Clifford, 1982; Pavesi, 1986; Schmidt, 1983).

### *Focus on Form*

The *Focus on Form (FoF)* approach that has gained prominence since the 1980s emerged in opposition to both FoFS and FoM and is sometimes referred to as the form-and-meaning focused instruction where learners are encouraged to communicate while paying attention to form (Purpura, 2004). Long (1997) considered FoF to be a viable third option that adequately captures the strengths of the analytic approach while at the same time attempting to overcome its limitations.

FoF constitutes a core methodological principle in Long's (1996, 1997, 2000) conception of task-based language teaching (TBLT) and refers to a variety of pedagogic procedures designed to shift the students' attention briefly to language features in the course of performing a classroom task where the focus is otherwise on meaning and successful achievement of a nonlinguistic real-world purpose. Ideally, this brief shifting of form should occur when the learners are experiencing a communicative need for learning the new grammatical structure, and, therefore, know exactly what they want to say but lack the adequate linguistic means to say it (Doughty & Williams, 1998). This latter consideration makes FoF drastically different from the traditional instruction, or FoFS, where the teacher introduces both the form and the meaning that frequently has to be explained in very abstract terms. For example, it is difficult to explain the meaning of such English structures as "should have done" or "could have done" outside a communicative use situation. In other words, the crucial distinction between FoF and FoFS is that FoF "entails a prerequisite engagement in meaning before attention to linguistic features can be expected to be effective" (Doughty & Williams, 1998, p. 3). Some researchers, however, notably Swan (2005), have objected to this perceived

“polarization” of meaning-based and form-based instruction that they believe to be a direct result of “damaging ideological swings” in language teaching theory and practice (p. 376).

To add to the controversy, some of the “founders” of the FoF approach, most notably Long (1996, 1997), did not recognize so-called *preemptive* or *planned* FoF. In contrast to other researchers, Long originally defined FoF as purely *reactive* in nature. This position appears to be quite extreme because the reality of most language classrooms with their intensive nature of instruction, planned tests, and absence of language-immersion environments outside of class (as in the case of FL vs. L2 learning) makes a proactive syllabus a necessity.

Many researchers and course designers have adopted Doughty and Williams’ (1998) definition of FoF, rather than Long’s (1996, 1997) original definition. Doughty and Williams’ definition included planned and proactive FoF activities as long as the following three major factors were taken into consideration: (a) the need for learners to be engaged in meaning prior to giving attention to the linguistic code used to express it, (b) the importance of identifying the learners’ actual language problems that require intervention, and (c) the need to keep the grammar intervention unobtrusive and fairly brief so they do not detract from meaning-focused activities. This approach is what Lightbown (2007), among others, has referred to as putting grammar instruction “in its proper place.” In his more recent writings, Long (2000) appeared to have adopted Doughty and Williams’ (1998) broader definition of FoF that also is used in this study.

In addition to FoF, another term that frequently is used in the SLA field to refer to instruction that focuses learner attention on linguistic features is *form-focused instruction*

(FFI). It is an umbrella term that covers “any planned or incidental instructional activity that is intended to induce language learners to pay attention to linguistic form” (R. Ellis, 2001, p. 1). FFI, in contrast to FoF, can refer to all techniques that draw the learner’s attention to grammatical structures (Williams, 2005). R. Ellis distinguished between three types of FFI: (a) FoFS, (b) planned FoF, and (c) incidental FoF. The present study was only concerned with the effectiveness of task-based interaction in focused oral-communication tasks as one of the planned FoF techniques.

*Task-based Interaction as a Focus-on-Form Instructional Technique*

Nassaji (1999) discussed two major ways to incorporate FoF into communicative activities in the classroom: (a) by process (i.e., incorporating learner- or teacher-initiated FoF naturally, as the need arises, throughout the instructional process), and (b) by design (i.e., designing these activities with a predetermined focus on specific forms). A number of L2 and FL educators have advocated a task-based approach to the teaching of grammar as the ideal way to accomplish a focus on form within meaningful, purposeful communication since the 1980s (e.g., Breen & Candlin, 1980; R. Ellis, 2003; Lee, 2000; Long & Crookes, 1993; Nunan, 1989; Van den Branden, 2006).

According to Larsen-Freeman (2001a), the principles of designing such activities are simple: the teacher finds an engaging activity with a clear real-world goal for the students’ interaction that will require them to produce the targeted grammatical structures repeatedly. In reality, however, classroom teachers frequently point out lack of time, lack of creativity, and difficulty of matching tasks to specific units of the course as challenges in coming up with successful focused communication tasks (Cobb & Lovick, 2007). The issue of creativity is a challenging one because poorly designed, unnecessarily complex,



or unmotivating tasks do not achieve their pedagogical goals. Creativity generally is understood as being able to present a product that is both interesting and original as well as feasible, practical, and well-suited for the end-user (Csikszentmihalyi, 1996; Sternberg, 1999). As Samuda (2007) pointed out, creating so-called pedagogic spaces where the learners can be successful with a contextualized task, frequently requires the efforts of a skilled curriculum designer. Regardless of the source of the task, the classroom teacher needs to be skilled in engaging the learners' attention and interest, monitoring their on-task behavior, providing both strategic and linguistic help when needed as well as appropriate feedback (Cobb & Lovick, 2008). Undoubtedly, some structures lend themselves better to focused task design than others. Additionally, learners can be quite adept at sidestepping the predesigned grammatical focus during task performance (R. Ellis, 2002), so priming for the target structure may need to be provided before the task commences.

There are some additional research-supported task-design considerations. Research findings have shown that learners are more successful with better-structured tasks, that is, more specific workplans, or instructions, lead to higher quality of the learners' TL output (Skehan & Foster, 2001; Willis, 1998). Additionally, because learners' attentional resources are limited (Skehan, 1998), they should not have to attend to too many novel elements at once during task performance (R. Ellis, 2005). The amount of attention that learners can allocate to grammar will be greater if other aspects of their performance are automatized or at least scaffolded (i.e., supported; R. Ellis, 2003). For example, in order to minimize the cognitive burden of lexical searches, such tasks can be

performed with well-known lexical material, or pretask help can be provided with the lexical as well as the pragmatic and sociocultural aspects of the task.

R. Ellis' (2003) conceptualization of FFI within CLT includes measured use of elements of FoFS such as overt preemptive grammar explanations and elements of drill-like activities when the teacher deems necessary. As pointed out earlier, in contrast to Long's strict view of TBLT where a task is the only viable unit of curriculum, instruction, and even assessment, R. Ellis (2003) emphasized so-called *task-supported* language teaching, that is, integration of task-based activities with more traditional elements of the curriculum. R. Ellis pointed out that there are opportunities for attention to form in all three phases of task performance: (a) through input flood, enhanced input (i.e., typographical enhancement of target structures in the input), modeling, and so forth in the pretask phase, (b) through appropriate corrective feedback in the during-task phase, and (c) through reflection, consciousness-raising, controlled-practice activities, and so forth in the posttask phase. When FoFS techniques are used in the pretask or posttask phase, the presence of these instructional elements that was not addressed in Keck et al.'s (2006) meta-analysis can become an important moderator variable influencing the effectiveness of the task itself as instructional treatment. In the present meta-analysis, the presence or absence of these elements was recorded on the coding forms when the information was available and presented clearly in the included primary studies.

In summary, focused communication tasks provide learners with opportunities to practice grammar through interaction while conveying personal meaning (unlike in mechanical drills) without sacrificing accuracy. When conducted with adequate scaffolding, guidance, and monitoring, these activities induce repeated use of target

structures under natural conditions of real-time communication, requiring the students to invest larger amounts of mental effort, thus contributing to deeper processing and better long-term retention. Varied, well-designed grammaring activities that are within the learners' zone of proximal development help keep up learner motivation and engage learners who may otherwise tune out during grammar practice. Additionally, unlike traditional types of practice, focused communication tasks constitute transfer-appropriate processing of target language structures, that is, processing that is conducive to developing skills transferrable to real use situations in TL speaking environments (DeKeyser, 2007). Effective teachers, however, typically subscribe to an eclectic approach to grammar teaching where they draw on a variety of instructional techniques depending on the goals of the program as well as the needs, cognitive styles, and inclinations of individual students (Purpura, 2004). Excessive reliance on one approach or recipe is unproductive and does not suit all target structures or learners. Integration of diverse and creative techniques, rather than polarization of extreme views, should be the core of language teacher education, particularly as it relates to the teaching of grammar.

#### *Types of Target Structure as Moderator Variables*

One of the variables that possibly moderates the effects of task-based interaction (Research Question 4) is the type of TL structure targeted by this instructional activity. It is not known whether task-based interaction in focused communication tasks is effective differentially for acquisition of different types of structures. A well-defined, comprehensive classification of target grammatical structures does not exist; however, attempts to classify structures along various dimensions are discussed below.

It frequently is pointed out in SLA research literature that not all grammatical items appear to be equal in terms of their learnability, that is, some are relatively straightforward and unambiguous for learning whereas others are ambiguous and complex, at least for students whose L1s are dissimilar to the TL (DeKeyser, 1998; Doughty & Varela, 1998; R. Ellis, 2006a; VanPatten, 1996, etc.). It has proven difficult, however, to pinpoint the exact features that make structures more or less learnable.

As presented in previous sections, the issue of learners' developmental readiness is involved (Pienemann, 1984, 1989). If the learners are not yet at a stage where the new structure is within their zone of proximal development (Vygotsky, 1986), no well-designed learning activities can help them acquire the structure. The true measure of successful acquisition of a target structure is the learner's demonstrated ability for spontaneous, relatively effortless, and errorless processing of the structure as it comes up in communication, rather than ability to produce the form or to provide the associated rule when prompted by the teacher (Nunan, 1999).

Individual learner differences play a role as well. Such potential moderator variables as learner age (Han, 2004, Seol, 2007), L1-L2 differences (Seol, 2007), learner aptitude (Fujii, 2005), and so-called orientation to form (Fujii, 2005) have been presented in the previous sections. Tomlinson (2007) who placed great emphasis on inductive discovery-learning of grammatical structures acknowledged that, when a student has a particular cognitive style that is not well suited for language analysis or when the rule is "convoluted," it makes more sense to present the rule deductively. Stating a rule explicitly may result in bringing about linguistic insights in a more efficient manner, as long as the teacher does not oversimplify the rule or present it in such a manner that the

students struggle more to understand the explanation than to apply the rule relying on their implicit knowledge of it (Larsen-Freeman, 2001b).

There are other factors that may interfere with the teachability of a particular structure that have to do with the characteristics of the structure itself. For example, Zhou's (1991) research findings demonstrated that formal instruction resulted in acquisition of some target structures (i.e., passives) but not others (i.e., tense and aspect). Jackson (2008) speculated about whether German accusative case-markings can be considered ambiguous or unambiguous. There is no clear answer to this question because masculine nouns in German are preceded by the case-marking information that clearly identifies the grammatical role of the noun in question, whereas case-markings for both feminine and neuter nouns are identical in their nominative and accusative case-forms (i.e., in the grammatical roles of the subject and the direct object).

Lee and VanPatten (2003) claimed that, due to the differences between structures, explicit explanations are not always necessary because some rules may be processed effectively by the learners without the teacher's assistance. Krashen (1981) argued that all really complex structures can only be acquired implicitly. Doughty and Varela (1998) acknowledged that some target structures may require little or no instruction, whereas other structures such as English articles are "remarkably impermeable" to it. Conversely, certain concepts appear to "beg" for direct instruction such as the verbal aspect in Russian (Leaver, 2000) because inferring the meaning of aspect from context may not be possible for learners, at least not without investing a lot of time and energy.

In terms of the role of instruction for grammar rules of various levels of difficulty, DeKeyser (2003) asserted that instruction is not really useful for very easy rules (because

it is not necessary) or for very complex rules (because it is not effective). The complexity of a grammatical item may be described in terms of its formal and functional complexity (DeKeyser, 1998). A structure may be complex formally when it requires complex processing operations, and it may be complex functionally if the relation between its form and its meaning is opaque (i.e., difficult to grasp). DeKeyser admitted that this classification is difficult to apply in practice and researchers disagree about how specific structures should be labeled using this classification. For example, Krashen (1981) deemed the English third-person singular -s ending formally simple, whereas R. Ellis (2006b) and DeKeyser (1998) considered it complex because it has to agree with the subject of the sentence and simultaneously denotes several semantic characteristics (i.e., the present tense, singular number, and third person). It has been well documented that large numbers of ESL and EFL learners fail to mark the third-person singular present even at more advanced learning stages even though they have been taught the rule many times (DeKeyser, 1998).

Hulstijn and De Graaf (1994) considered the target structures in view of how many different criteria, or linguistic transformation rules, the learner needs to apply in order to arrive at the correct target form. Spada and Tomita (2010) used this classification in their meta-analysis of the effects of explicit over implicit instruction for simple and complex target grammatical features in English. For example, according to Spada and Tomita, English *wh-questions* where the question word functions as object of a preposition require seven transformations to arrive at a sentence like “Who did you talk to”? whereas past tense of regular English verbs requires only one transformation (i.e., addition of the *-ed* inflection). The meta-analysts decided to code target structures that

require only one transformation as simple and those that require two or more transformations as complex.

Additionally, Hulstijn and De Graaf (1994) identified such factors as *scope* (i.e., absolute number of instances in which the target structure appears) and *reliability* (i.e., percentage of instances under which the associated rule holds) as having an effect on acquisition of a target structure. These factors relate to whether learning of specific structures is primarily rule-based or exemplar-based. Hulstijn and De Graaf hypothesized that explicit instruction is beneficial only for structures governed by rules of large scope and high reliability. In the case of a structure of limited scope, learning of the rule may not justify the effort. In the case of an “unreliable” rule, learners may benefit more from exemplar-based learning of the target structure because the associated rule cannot be applied safely. For example, due to the complex nature of morphological variation in inflecting-fusional languages (e.g., Russian) learners of these languages may need to engage in so-called exemplar- or item-based learning for many grammatical structures (Kempe & Brooks, 2008).

A more comprehensive classification was offered by R. Ellis (2006b) who attempted to make a distinction between the structure’s difficulty from the point of view of, on the one hand, implicit knowledge and, on the other hand, explicit knowledge. He classified grammatical structures from the point of view of implicit knowledge based on the following criteria: (a) frequency, (b) saliency (i.e., how easy it is to notice a particular form in TL input), (c) functional value (i.e., whether the form maps onto a clear and distinct function), (d) regularity (i.e., whether the form conforms to a clear, identifiable pattern), and (e) processability (i.e., how easy it is for the learner to process the form).

Regarding the explicit knowledge required for mastery of a structure, according to R. Ellis, it may be hypothesized as the degree of conceptual clarity and the degree of difficulty of the meta-language involved in the explanation of a particular structure. R. Ellis' research findings showed that items that were not difficult from the point of view of implicit knowledge were often difficult in terms of explicit knowledge and vice versa. A regression analysis, however, demonstrated that both kinds of knowledge predict general language proficiency. R. Ellis speculated that it would appear that learning of some structures depends on general acquisition processes, whereas learning of others is based on something more akin to problem-solving ability. As general language proficiency appears to be enhanced by both types of knowledge (i.e., explicit and implicit), it is important to draw on both types of knowledge in classroom teaching practices.

Keck et al. (2006) pointed out the need to investigate the possible moderating effects of the type of target structure on the effectiveness of task-based interaction as an instructional treatment. Considering the vast and intricate differences that exist between the world languages (MacWhinney, 1995), it is not easy to apply any of the classifications presented so far across languages. Primary researchers sometimes have attempted to categorize the target structures in their study reports. For example, in his influential study, Robinson (1996) labeled so-called pseudo-clefts of location in English a "hard" rule for Japanese learners, whereas subject-verb inversion following adverbial fronting was labeled an "easy" rule. The Spanish contrary-to-fact conditional forms were labeled a "complex" structure in Rosa and O'Neill's (1999) study. If the primary



researcher does not provide such information, making such distinctions between target structures is a challenging undertaking for the meta-analyst.

The challenge lies in the fact that languages differ significantly in the extent of nominal-declension and verbal-conjugation forms as well as in ways in which such grammatical characteristics as tense, space, number, and so forth are marked (Bloomfield, 1961; Greenberg, 1978). Besides learning the formal word classes of the TL, in some languages, learners are faced with having to learn new paradigms of thinking that may be very different or nonexistent in their L1 as, for example, in the case of speakers of English learning Korean or Japanese (MacWhinney, 1995). Even such a fundamental rule as the one distinguishing between the use of *ser* and *estar* (both meaning “to be”) in Spanish is an example of new conceptual understanding required of the learners in addition to knowledge of all the conjugated forms of these two verbs.

To sum up, classifying the types of target structures consistently across many TLs, most of which may not be known to the meta-analyst, is a truly challenging undertaking. Nevertheless, every effort was made by the meta-analyst in the present study to record all data pertaining to the nature of the target structure(s) that are reported or can be inferred from the primary study reports. As a minimum, the meta-analyst and the second coder recorded whether the structure was morphological or syntactic in nature, whether it was simple or complex based on the classification used in Spada and Tomita’s (2010) meta-analysis adopted from Hulstijn and De Graaf (1994), and whether it appeared to be ambiguous or unambiguous conceptually.

#### *Degree of Task Essentialness of the Target Structure*

Of great interest to researchers is the possibility of injecting, or “seeding,”

specific targeted language items into a task without compromising the communicative nature of the task (Skehan & Foster, 2001). Loschky and Bley-Vroman (1993) and Samuda, Gass, and Rounds (1996), among others, suggested that tasks can be designed so as to make the use of specific target items by the learners highly probable, if not unavoidable. Others (Skehan & Foster, 2001; Willis, 1996) were critical of this assertion pointing out that it lacks strong empirical evidence, especially, of the feasibility of application of such task designs to a wide range of target grammatical structures.

Loschky and Bley-Vroman (1993) distinguished three ways in which a task can be designed to target a specific language structure:

1. *Task naturalness*. The targeted grammatical structure may not be absolutely necessary to complete the given task; however, the need for it may arise quite naturally when learners interact to complete the task and, therefore, they are likely to use it (R. Ellis, 2003; Keck et al., 2006; Loschky & Bley-Vroman, 1993). For example, Tarone and Parrish (1988) found that a narrative task is likely to elicit use of definite noun phrases, whereas an interview task is more likely to elicit use of indefinite noun phrases.

2. *Task utility*. Although the targeted structure is not absolutely essential for completing the task, it is very useful (R. Ellis, 2003; Keck et al., 2006; Loschky & Bley-Vroman, 1993). For example, a spot-the-difference picture task where differences involve spatial relations can be performed without the use of certain prepositions and adverbial phrases (e.g., “by,” “next to,” “to the right of,” “behind,” etc.) however, the knowledge of these items is very useful for completing such a task.

3. *Task essentialness*. In this case, the learners must use the target structure to achieve a satisfactory outcome (R. Ellis, 2003; Keck et al., 2006; Loschky & Bley-

Vroman, 1993). A participant's incorrect use, avoidance, or inability to understand the target item renders completion of the task by the group impossible. Admittedly, although it is easy to design comprehension-based tasks in which knowledge of a particular structure is required, it is much more challenging to design a production-based task in which it would not be possible for learners to circumvent the use of the target structure somehow (R. Ellis, 2003; Loschky & Bley-Vroman, 1993). Arguably, learners can be directed to use the target structure (Lightbown, 2007), however, some researchers (Skehan, 1998) insisted that it invalidates the task purpose.

Some primary researchers have included evidence of the degree to which the learners actually used the target structure during task completion in their study reports in the form of transcripts or usage counts (Mackey, 1999; Tuz, 1993). It has been reported that eliciting the use of some target structures through task design is easier than eliciting others. For example, eliciting question forms (Mackey, 1999) appears to be fairly easy, whereas eliciting noun phrases with multiple attributive adjectives that have to be used in the appropriate order is much more elusive (Tuz, 1993).

In their meta-analysis, Keck et al. (2006) used Loschky and Bley-Vroman's (1993) definitions for coding of the task-essentialness, task-usefulness, and task-naturalness of the target structure relevant to the treatment task. The meta-analysts reported that, on immediate posttests, tasks with task-essential target structures ( $d = .83$ ) were found to have somewhat smaller effects than those with task-useful target structures ( $d = .98$ ); however, the 95% confidence intervals were overlapping. A larger difference with nonoverlapping confidence intervals was observed on so-called short-delayed posttests up to 29 days after the interaction treatment took place. Keck et al.'s meta-

analytic findings supported the claims made earlier by Doughty and Varela (1998), Loschky (1994), and Loschky and Bley-Vroman (1993), among others, that task-essentialness is an important moderating variable in learner acquisition of the target structures.

In the present meta-analysis, the task-essentialness of the target structure was coded and examined as a moderator variable. If the primary researcher did not report the degree of task-essentialness, the two coders used their best judgment to make an inference decision as Keck et al. (2006) did in their meta-analysis. Evidence based on transcripts and target structure usage counts was examined to aid in making such decisions when available. Table 3 presents a summary of all variables that may affect learners' success with target structure acquisition through task-based interaction in focused oral-communication tasks that have been discussed so far in chapter II. Because the range of the variables potentially moderating the degree to which the learners benefit from task-based interaction is wide, the list of presented variables cannot be considered exhaustive. Some of the variables presented in Table 3 have not been discussed in detail in the earlier sections due to the fact that they were not represented sufficiently in the primary study reports included in this meta-analysis.

The majority of these variables and their levels are represented in the coding form (see Appendix C). Information about others, when available, was entered under "additional information" in the coding form. The next section presents a discussion of measurement issues as related to the acquisition of the target structures and the potential effects of the type of outcome measures on primary study outcomes.

Table 3

Summary of Variables Potentially Affecting Learner Acquisition of the Target Structure Through Task-Based Interaction

Task-Related Variables	Learner-Related Variables	Variables Related to Target Structure	Pedagogical Variables	Other Variables
Task design (information-gap, reasoning-gap, etc.)*	Age, gender, cultural background, L1	Morphological vs. syntactic	Presence of explicit instruction (rule explanation, modeling, etc.)	Teacher's familiarity with TBLT
Task open-endedness*	Cognitive characteristics	Complexity	Presence of error correction	Teacher's metalinguistic knowledge
Task convergence*	Proficiency level	Ambiguity	Presence of pretask planning	Teacher's attitudes to TBLT
Task complexity	Personality traits (active vs. passive, etc.)	Task-essentialness	Characteristics of the interlocutor	FL vs. L2 context
Task difficulty	Personal goals, motivation, etc.		Presence of additional instructional elements (input processing, traditional drills, etc.)	Institutional expectations
Task origin	Familiarity with TBLT Attitudes toward TBLT			Type of outcome measure**

\* These types of tasks are presented in more detail in Table 2.

\*\* This variable is presented in the next section.

### Measures of Acquisition of Target Grammatical Structures

The dependent variable in the primary studies that were included in the meta-analysis is the learning, or acquisition, of the target structures as measured by the students' scores on immediate, short-delay, or long-delay posttests. The operationalization of this variable, common measurement instruments, and issues associated with these instruments are presented in this section.

One of the main issues in measuring learners' acquisition of target structures is the predominant use of traditional, grammar-translation measures that are not in line with communicative teaching (Gass & Mackey, 2007; Norris & Ortega, 2000). Traditional approaches to testing FL and L2 learners' mastery of specific grammatical items usually entail discrete-point, decontextualized language use (Larsen-Freeman, 2003). For example, a typical item on a grammar test may require learners to fill in the blanks with the correct endings or to select the correct language form among several forms provided. Students also traditionally have been asked to make judgments about the grammatical correctness of a sentence or to provide the associated rule (Larsen-Freeman & Long, 1991). Such testing formats have been criticized extensively both from the pedagogical and from the research perspectives (Gass & Mackey, 2007). In discussing various data-collection measures used in second and foreign language acquisition research, Gass and Mackey asserted that research findings are "highly dependent on the data collection measures used" (p. 4). Although SLA research largely is focused on investigating the effectiveness of various explicit versus implicit methods of teaching grammar, traditional outcome measures undoubtedly favor explicit treatments (Norris & Ortega, 2000) because their format is consistent with traditional explicit grammar instruction. It has been argued that research outcomes may be test-dependent, that is, learners who acquired a grammatical structure implicitly will perform better on measures of implicit knowledge, whereas those who learned it under conditions of explicit instruction will perform better on grammaticality judgment and other similar measures of explicit knowledge (Erlam, 2003). If teachers use primarily traditional formats for testing grammar, it is reasonable to assume that students who have been taught grammar mostly through communication,

which is consistent with the current communicative approach, may be at a disadvantage when it comes to traditional grammar tests as compared with their peers who have been taught in a more didactic, noncommunicative manner.

On a related note, a major debate still remains on the research agenda due to a lack of consensus of how acquisition of grammar can be measured, that is, specifically what types of assessment tasks make it possible to infer that grammatical knowledge has been acquired (Purpura, 2004). For example, Mackey (1999) pointed out the following methodological challenges that face researchers who conduct empirical explorations of the relationship between conversational interaction and development of mastery of grammatical forms: (a) difficulties in designing tests that would measure acquisition of target forms used during the interaction and (b) difficulties in operationalizing acquisition, or development, of these target forms in the first place.

In line with the current communicative language teaching methodology, it is important to move from traditional ways of assessing knowledge of grammar toward assessing mastery of its use through communicative tasks (Norris & Ortega, 2000). Such assignments, unlike discrete-point exercises, resemble real-world tasks that learners eventually will have to complete in real-time interaction with others. Otherwise, a learner hypothetically may do well on a discrete-point grammatical test but be unable to demonstrate true grammatical competence while performing in real, spontaneous interactions due to the so-called inert knowledge problem (Larsen-Freeman, 2001a). Unfortunately, the use of communicative tasks as outcome measures is still limited both in research and, in particular, in classroom assessment (Gass & Mackey, 2007). Common

techniques that currently are used in outcome measures designed for assessing the students' grammatical competence are presented in the subsequent subsections.

### *Common Data-collection Techniques in Outcome Measures*

This subsection presents a description of common testing formats used to measure L2 acquisition. It mainly focuses on elicitation techniques that collect data for measuring (a) production of the target structure, (b) comprehension of the target structure, and (c) metalinguistic knowledge about the target structure.

### *Naturalistic versus Elicited Data-collection Procedures*

Chaudron (2003) distinguished between naturalistic observations of the participants' use of TL during play, normal interactions, and classroom interactions, on the one hand, and so-called elicited production procedures, on the other hand. Although naturalistic observations can render valuable data about a learner's interlanguage grammar development, the use of this technique is costly, time-consuming, and, if more than one participant from the group is involved, challenging. Moreover, Chaudron warned that, if teachers or researchers are concerned with the development of a specific target structure, they may have to wait for a long time for this structure to appear with sufficient frequency in the learner's speech sample during truly unstructured naturalistic observations. For this reason, teachers and researchers employ a number of elicitation techniques that allow them to measure the development of specific language items in the learners' interlanguage.

### *Elicitation of Production Data*

Based on Bialystock's (1988, 1994) cognitive model of SLA, researchers typically measure acquisition along two cognitive processing components: (a) analysis of



knowledge and (b) control of processing. Larsen-Freeman and Long (1991) provided a list of the following instruments used to collect language production data (pp. 27-30):

1. Reading aloud (typically when the focus is on pronunciation).
2. Structured exercises requiring some sort of grammatical manipulations such as transformation exercises, fill-in-the-blanks, sentence-rewrite, sentence-combining, and multiple-choice measures.
3. Completion exercises such as completing a sentence when only the beginning or the end is provided or completing the missing parts of a dialogue.
4. Elicited imitation wherein the participants are asked to repeat or imitate utterances that are too long to be held in their short-term memory (thus forcing them to rely on their understanding of the morphosyntax or even on using their own TL grammar to construct the imitations).
5. Elicited translation (typically from L1 into TL).
6. Guided composition (e.g., based on picture sequences, a list of content words, or a silent video).
7. Questions and answers (typically based on a single picture, a series of pictures, or a prescribed personalized situation).
8. Reconstruction (i.e., oral or written story retelling based on a printed text, an audio, or a video).
9. Communication games (e.g., finding out from a partner how objects are arranged in a picture in order to imitate this arrangement or describing a picture so that the partner can figure out which one from a set is being described, etc.)

10. Role plays (particularly useful in eliciting data about the use of honorifics and other forms appropriate when talking to an interlocutor of a particular status in some languages).

11. Oral interviews in which the interviewer may ask questions that are likely to elicit the use of specific forms (e.g., asking participants what they did over the weekend will most likely result in the use of past tense forms).

12. Free composition.

These elicitation measures are arranged in the order of increasing degree of control that the learner (vs. the test designer) has over the choice of language items to be used. In terms of communicative games and tasks, Gass and Selinker (2008) listed the following additional examples of measures that can be used for eliciting L2 speech samples containing specific grammatical items: (a) picture descriptions (e.g., when the pictures contain details that predispose learners to using specific language items, for example, prepositions and adverbs of location), (b) tasks requiring “spotting the differences” between two pictures (e.g., when the differences intentionally are created in such a way that learners are likely to use specific language items when talking about them), (c) consensus-reaching tasks, and so forth.

### *Elicitation of Comprehension Data*

Some outcome measures are designed so as to obtain evidence only of participants’ comprehension (vs. production) of specific target structures. Among these measures are so-called truth-value judgments (Gass & Selinker, 2008) or sentence verification measures (Loschky, 1994) that are frequently used to test understanding of reflexive pronouns in English (e.g., “Mr. Smith saw him” vs. “Mr. Smith saw himself”).

In such measures, the participants have to state whether the given sentences are true to fact based on a picture, a series of pictures, or a presented situation. Other examples of comprehension measures include translation from TL into L1 (Larsen-Freeman & Long, 1991), following directions (Gass & Varonis, 1994), or so-called “act-out” activities (Chaudron, 2003) in which the participants have to respond by performing actions, which is only possible if they are able to understand instructions containing targeted structures adequately.

#### *Elicitation of Metalinguistic Data*

Larsen-Freeman and Long (1991) listed the following procedures that traditionally are used to elicit knowledge of metalinguistic information about targeted language items, or the so-called “intuitional data elicitation” instruments (pp. 33-34):

1. Error recognition and correction (i.e., participants are asked to recognize and correct errors in their own utterances or utterances produced by other learners),
2. Grammaticality judgment (i.e., judgment about whether or not a particular utterance is well-formed grammatically),
3. Other judgment measures (e.g., making judgments about the relative politeness or formality of a particular utterance), and
4. Card sorting (i.e., categorizing or ranking sentences presented on cards based on some principle, for example, based on whether male or female gender forms are present).

R. Ellis (2006b) pointed out that metalinguistic grammaticality-judgment tests can be timed or untimed. Typically, timed grammaticality-judgment tests are administered via computer.

### *Types of Outcome Measures as Moderator Variables*

This subsection presents the classification of outcome measures that was used in the present meta-analysis to investigate the mediating effects of outcome measures on study outcomes in conjunction with Research Question 5. Norris and Ortega (2000) who conducted a research synthesis and meta-analysis of empirical studies investigating the effectiveness of different types of L2 grammar instruction reported that individual researchers employed a variety of different outcome measures that ranged from discrete-point tests requiring a mere display of grammatical knowledge to free oral production in which the participants conveyed personalized meaning. The meta-analysts created a classification for coding these diverse outcome measures across studies. Outcome measures were coded as *metalinguistic grammaticality judgment* if examinees are required to evaluate the grammatical correctness of utterances containing the target structure in its appropriate target-like form or inappropriate nontarget-like form. *Selected-response* measures entailed participants choosing the grammatically appropriate item containing the target structure out of the ones provided. Instruments that required examinees to produce TL segments ranging from a word form to a full sentence such as in items requiring substitution, transformation, sentence combining, or answering a simple question were coded as *constrained-constructed response*. Instruments requiring participants to produce more extended monologic discourse, whether written or oral (e.g., written composition, oral story based on pictures, etc.) were coded as *free-constructed response*.

Norris and Ortega's (2000) classification was adopted for the present meta-analysis; however, an additional category was added to reflect the growing trend in

primary research. This category was labeled *oral-communication task* and was used in situations when the outcome measure itself represented an interactive, communicative activity that met the definition of a task provided earlier in this chapter. Typically, when this category of outcome measure was used, it was an oral-communication task similar to the task(s) used as the treatment in the study.

One of the research questions in Norris and Ortega's (2000) meta-analysis was whether the type of outcome measure influences observed instructional effectiveness. Their findings suggested that the type of outcome measure likely affects the magnitude of the observed effect. Specifically, observed effects were likely to be greater if the outcome measure involved selected-response or constrained-constructed-response formats, and smaller if the outcome measure involved metalinguistic-judgment or free-response formats. In general, however, only 10% of all tests used in the primary studies included in Norris and Ortega's meta-analysis used some sort of a free-response measure. For the purposes of the present meta-analysis, use of noncommunicative tests to measure acquisition of grammar through task-based interaction potentially raises testing validity issues. The next section addresses this and other issues in measuring acquisition of L2 and FL grammatical structures.

#### *Issues in Measuring Acquisition of Grammatical Structures*

This subsection provides an in-depth presentation of such issues as limitations of discrete-point outcome measures and lack of standardization and reporting consistency. This presentation is important for situating the discussion of effects of the type of outcome measure as a potential moderator variable.

Among the acquisition measures listed in *Common Data-collection Techniques in Outcome Measures*, the following perhaps are associated most closely with traditional approaches to classroom testing of grammar: (a) structured exercises (i.e., fill-in-the-blanks, selecting the correct grammar form, etc.) and (b) elicited translation. Classroom instructors sometimes also use reading aloud at beginning stages; however, the purpose is typically to test letter-sound correspondence and pronunciation rather than grammar. Although it is possible that grammaticality judgments and other intuitional data-collection methods are used in classroom tests, these outcome measures perhaps are most likely to be encountered in research studies due to the fact that they offer insights into what utterances the participants consider to be appropriate and native-like. In order to ensure that the fact of judging an utterance to be ungrammatical is not based on a mere guess or on the evaluation of a feature that is not related to the grammatical structure in question, some experimenters require the participants to correct all utterances deemed to be ungrammatical (Gass & Mackey, 2007).

Due to the fact that it is easier to develop, standardize, and calculate reliability for discrete-point tests, it is not surprising that teachers and researchers use discrete-point tests most frequently (Chaudron, 2003). Additionally, Mackey (1999) suggested that some target structures are easier to elicit than others under testing conditions. She explained that her choice of question forms as the targeted structure for her experimental study was based on the fact that previous research findings indicated that question forms could be “readily elicited” (p. 566).

Based on their classification presented in the previous section, Norris and Ortega (2000) reported that approximately 90% of study outcome measures utilized discrete

routines (i.e., meta-linguistic judgments, selected responses, constrained-constructed responses), and a mere 10% involved extended communicative use of the TL (i.e., free-constructed responses). This is an unfortunate finding because the use of communicative task-based interaction in the teaching of grammar has been demonstrated to result in larger effect sizes than teaching through activities not requiring such interaction (Keck et al., 2006). Additionally, as stated earlier, there appears to be a correlation between the participants' scores on outcome measures and the congruency of these measures with the instructional methods used (i.e., learners who have been taught grammar communicatively are, on average, expected to do better on communicative measures and vice versa; Erlam, 2003; Gass & Mackey, 2007).

In discussing research on the effectiveness of corrective recasts as a feedback technique, Long (2007) pointed out that the issue of reliability and validity of outcome measures largely is ignored in the literature. Norris and Ortega (2000) reported that only 16% of the reviewed studies attempted to report any information related to reliability or validity of the outcome measures. Additionally, the primary studies varied widely in the extent to which targeted language forms were tested by outcome measures. Norris and Ortega reported that some studies utilized only one test item per targeted structure, whereas others employed lengthy tests with multiple items per structure or elicited extensive language production data. The number of dependent variables varied between one and four in any single study.

To complicate matters further, according to Norris and Ortega (2000), individual researchers employed different techniques in evaluating the responses that participants provided on outcome measures: (a) dichotomous measures (i.e., correct or incorrect), (b)

polytomous measures (e.g., subjective ratings of relative appropriateness), (c) measures based on error frequency counts, and (d) measures based on identified stages in interlanguage development (rarely used due to the challenges of identifying the stages). (The scoring procedures used in the primary studies included in the present meta-analysis also were very diverse.)

Mackey's (1999) operationalization of development may serve as an example for illustrating rarely used stages-based evaluation measures. She operationalized development, or acquisition of the target structures (question forms), as the learners' progression, or lack thereof, through the sequence previously identified for English question formation by Pienemann and Johnston (1987). This progression typically involves movement from incorrect canonical word order (e.g., "I can draw a house here?") through several other stages toward correct inverted word order (e.g., "Can I draw a house here?"). It also involves mastery of structural nuances such as lack of inversion in relative clauses (e.g., "Who bought a cat?"), appropriate use of question tags (e.g., "He bought a cat, didn't he?"), and so on. In Mackey's (1999) study, if the learners demonstrated production of forms typical of a particular stage, they were believed to be at that stage in their acquisition of question forms.

Additionally, some researchers such as Lyster and Ranta (1997) who investigated the effectiveness of corrective feedback used immediate learner production as a measure of learner uptake (i.e., ability to incorporate corrected forms into learner's own output). Others, like Mackey (1999) in her seminal study on the same subject, used delayed posttests. In view of such great diversity of the outcome measures used, it is not



surprising that researchers sometimes report very different results for the effectiveness of the same or similar instructional treatments.

From the point of view of the effect that testing practices have on classroom instruction, if testing within a communicative course is conducted through traditional noncommunicative assessment measures, then so-called negative washback effect takes place (Brown & Hudson, 1998). Washback refers to the natural tendency of teachers and students to tailor both the format and the content of learning activities toward upcoming tests (Bailey, 1996). A positive effect naturally will occur when testing procedures correspond to the course goals and objectives (Brown & Hudson, 1998). For example, the use of authentic texts and tasks in tests will generate beneficial washback (Bailey, 1996) because it is likely to cause teachers to use authentic materials and task-based activities in the classroom. Conversely, when tests use obsolete grammar-translation methodology, the communicative orientation in classroom instruction will suffer due to the negative washback effect of testing practices on teaching practices.

In SLA research, the outcome measures typically are connected to the theoretical framework under which the research is conducted (Gass & Mackey, 2007). For example, a researcher interested in the effectiveness of explicit grammar teaching likely will choose outcome measures that elicit evidence of the students' explicit knowledge about the target structure. Because the choice of the outcome measure tends to have an effect on research findings, it is important that a variety of measures be used for a given domain. Gass and Mackey warned that this recommendation should not be understood to imply that all data-collection methods are good equally and that the choice of a particular measure should be made in correspondence with the research questions. Clearly, a testing

instrument that requires students to fill in the blanks with correct grammatical endings does not necessarily provide reliable data about these students' ability to use the associated grammatical forms correctly and appropriately in oral task-based interaction when their attention is on meaning and not on form. The use of a well-designed communicative task that predisposes students to using these particular grammatical forms as an outcome measure will contribute to greater construct validity if the researcher is interested in measuring students' ability to use these forms in communication. Additionally, the choice of specific measures is affected by whether the researcher is interested in gathering evidence about the learners' ability to comprehend the target structure, to produce it, or both (Gass & Mackey, 2007; Larsen-Freeman & Long, 1991).

It is important to review and question the elicitation methods used both for regular tests administered as part of FL and L2 courses and in empirical research studies. According to Gass and Mackey (2007), it may be difficult to capture the phenomenon under investigation with only one outcome measure. Therefore, triangulating from multiple measures should be used as much as possible (Chaudron, 2003). For example, based on Bialystock's (1988, 1994) cognitive model of SLA, researchers may use explicit structural exercises or metalinguistic measures for the purposes of knowledge analysis and, at the same time, use elicited imitation and communicative tasks to gather evidence about the degree of control of processing.

Norris and Ortega (2000) recommended that primary researchers always consider the validity of dependent variables in terms of what kinds of interpretations can be based on them as well as estimate and report the reliability of the use of outcome measures. It would be naïve, however, to assume that use of communicative tasks for assessment does

not present its own imminent challenges. Implementing performance-based assessment in general poses some important challenges in task design, scoring, training of raters, feasibility, efficiency and cost effectiveness, reliability and validity, and so on (Johnson, Penny & Gordon, 2009; Lane & Stone, 2006). Task-based assessment in language learning presents these issues as well. In discussing a hypothetical example of a researcher investigating acquisition of passive forms by English-speaking learners of Japanese, Gass and Mackey (2007) pointed out that even well-designed tasks may fail to elicit use of the target structure due to learner avoidance or other reasons. In empirical research, it is a common practice to field-test task prompts by obtaining samples of native speaker responses as evidence that the use of the target structure is natural in performing the task set up by a particular prompt (Gass & Selinker, 2008). Sometimes researchers capture the interaction between learners by means of audiotaping and then transcribing the TL output produced during task performance (Gass & Mackey, 2007). Similar steps may be taken to ensure validity of regular classroom testing measures.

Continued efforts are needed in identifying techniques for designing language performance assessments and scoring procedures, as well as more research into the reliability and validity issues of task-based assessment. In the meantime, primary researchers may benefit from using several assessment measures of different types to gather adequate evidence of the learners' mastery of the same target structure. Table 4 summarizes the types of outcome measures presented in the preceding section and coded in the present meta-analysis as well as their congruence (or lack thereof) with CLT.

Additionally, all testing measures utilized in the included primary studies were classified as immediate and delayed. In case of a delayed posttest, the length of delay

Table 4  
Summary of Types of Outcome Measures

Type of Outcome Measure	Congruence with CLT
1. Metalinguistic judgment	No
2. Selected response	No
3. Constrained-constructed response	No
4. Free-constructed response	Yes
5. Oral-communication task	Yes

between the instructional treatment and the test was recorded in the coding form (see Appendix C) as well. The next section presents a detailed overview of the meta-analysis of task-based interaction by Keck et al. (2006) that is related most closely to the research topic of the present meta-analysis.

#### Review of Keck, Iberri-Shea, Tracy-Ventura, and Wa-Mbaleka's (2006) Meta-Analysis: Investigating the Empirical Link Between Task-Based Interaction and Acquisition

This section offers a detailed review of Keck et al.'s (2006) meta-analysis because it is related closely to the purpose of the present meta-analysis, even though there were considerable differences in the scope of the search of primary research literature, the search procedures, the definitions of some key constructs, and the potential moderator variables that were examined between Keck et al.'s (2006) meta-analysis and the present study. The purpose of Keck et al.'s meta-analysis was to synthesize the findings of all experimental and quasi-experimental task-based interaction studies published between 1980 and 2003 where the dependent variable was learners' acquisition of specific grammatical or lexical items. The meta-analysts reported that results from 14 unique sample studies showed that treatment groups substantially outperformed control and comparison groups in the acquisition of both grammar and lexis on immediate and delayed posttests.

Keck et al. (2006) investigated whether and to what extent task-based classroom interaction (i.e., conversational interaction in the TL that takes place among NNSs or NSs and NNSs in pairs or small groups while completing assigned oral communication tasks) promotes TL acquisition. The meta-analysts wanted to know whether there is a direct link (vs. merely an indirect link) between learners' interaction in classroom tasks and increased knowledge of specific TL items (both grammatical and lexical) if the tasks are designed in such a manner that they predispose the learners to using these target items repeatedly. Additionally, the meta-analysts investigated what task design features (e.g., so-called task-essentialness of the target language item) contribute to greater gains in acquisition of the target item. Therefore, Keck et al.'s meta-analysis was focused on the following research questions:

1. Compared to tasks with little or no interaction, how effective is task-based interaction in promoting the acquisition of grammatical and lexical features?
2. Is the effectiveness of interaction tasks related to whether the target feature is grammatical or lexical?
3. Are certain task types (e.g., information-gap) more effective than others in promoting acquisition?
4. How long does the effect of task-based interaction last?
5. To what extent do the following task design features impact the extent to which interaction tasks promote acquisition: (a) the degree of task-essentialness of target features and (b) opportunities for pushed output? (p. 95)

The target population for Keck et al.'s (2006) meta-analysis were adolescents and adults (i.e., age of 13 years and over) engaged in FL or L2 study. The meta-analysts explained that, because it is unclear how age affects task-based interaction processes, including studies that involve children under 13 years of age would have complicated the issue by introducing another variable into the analysis. The research domain was defined as all experimental or quasi-experimental task-based interaction studies published

between 1980 and 2003. In 1980, Long (1981, 1996) first proposed his interaction hypothesis that posited that interaction played a crucial role in the development of the learners' interlanguage systems. In the 1980s, Long and others also first defined the role of TBLT in developing the learners' control over the grammatical form (Long, 1981, 1985, 1989).

Studies were selected from Education Resources Information Center (ERIC), Linguistic and Language Behavior Abstracts, PsychInfo, and Academic Search Premier databases. Search terms included combinations of the following keywords: *interaction, negotiation, feedback, communicative, input, output, intake, uptake, review of the literature, empirical, results, and second language acquisition (and learning)*. Keck et al. (2006) also conducted both manual and electronic searches of nine journals in the SLA field: *Applied Linguistics, Applied Psycholinguistics, Canadian Modern Language Review, Language Learning, Language Teaching Research, Modern Language Journal, Second Language Research, Studies in Second Language Acquisition, and TESOL Quarterly*. Additionally, the meta-analysts reviewed three comprehensive SLA textbooks looking for potential candidate studies and review articles (R. Ellis, 1994a; Larsen-Freeman & Long, 1991; Mitchell & Myles, 1998).

The described search procedure originally identified over 100 studies. The number of qualifying studies was later reduced from 100 to 13 studies based on the inclusion and exclusion criteria. The following are the inclusion criteria outlined by Keck et al. (2006):

1. The study was published between 1980 and 2003.

2. The study measured acquisition of an FL or L2 by adolescents or adults (i.e., participants over 13 years of age).

3. The study utilized communication tasks that were used for the following purposes: (a) as the treatment of the study or (b) to create contexts for the application of the actual treatment under investigation (e.g., recasts used for the purposes of error correction).

4. The tasks used in the study were face-to-face dyadic or face-to-face group oral communication tasks.

5. The task(s) was or were designed to foster acquisition of specific grammatical or lexical features.

6. The study was experimental or quasi-experimental in design and either (a) measured gains made by one group after the treatment using a pre- and posttest design or (b) compared gains made by the treatment groups with those made by the control or comparison groups.

7. The report adequately described the tasks employed in the study so that these tasks could be coded for task characteristics.

8. The dependent variable(s), that is, posttest scores or gain scores, measured the acquisition of specific grammatical or lexical structures targeted by the treatment.

Studies that utilized descriptive or correlational designs, involved computer-based interaction tasks (vs. face-to-face oral tasks) as well as studies in which treatments did not target acquisition of specific grammatical or lexical items or where participants received additional treatments (e.g., written corrective feedback) were excluded. The 13

study reports that met all of the specified inclusion and exclusion criteria contained 14 unique study samples that contributed effect sizes to Keck et al.'s (2006) meta-analysis.

Keck et al. (2006) explained that they had decided not to combine within-study effect sizes even though Lipsey and Wilson (2001) recommended doing it in order to avoid the problem of nonindependence of effect-size values. The meta-analysts explained their decision by the need to be able to analyze information about how the characteristics of each task and each target TL linguistic feature impact the effect of the treatment. This analysis would not have been possible if the within-study effect sizes were combined for different types of tasks or different types of target linguistic features. The meta-analysts explained that, for studies that compared multiple treatments, separate effect sizes need to be calculated for each treatment. Similarly, if the study investigated effects for different TL features, separate effect sizes need to be calculated for each feature. Keck et al.'s recommendation were followed in the present meta-analysis.

Included studies were coded for both substantive and methodological features. Coded substantive features were established on the basis of the review of relevant literature and included task type (i.e., jigsaw, information-gap, problem-solving, decision-making, opinion-exchange, or narrative), degree of task-essentialness (i.e., task-essential, task-useful, or task-natural), and opportunity for pushed output (i.e., presence or absence thereof). The methodological features captured by the meta-analysts included various research design and reporting features (i.e., group assignment, type of the learners' language-proficiency assessment, and type of dependent measure), learner characteristics (i.e., L1 and TL proficiency level), characteristics of the treatment setting (i.e., educational setting) as well as information about the statistical procedures used, for



example, analysis of variance (ANOVA) or multivariate analysis of variance (MANOVA), and statistics reported (i.e., a priori alpha, exact  $p$ , inferential statistics table, strength of association, standard error, confidence intervals, and effect size).

Two of the researchers coded all 14 studies independently with an overall agreement ratio of .88 (Cohen's kappa was .77). Task-essentialness was determined to be a high-inference variable for the purposes of coding because, in absence of the transcripts of the actual learner interaction, it was hard to determine to what degree a particular target item was used by the learners during task completion. Therefore, in order to code for this variable, the researchers carefully considered the target item against the design of the task. If a conclusion was made that the task was expected to elicit the use of the target item *by design*, then the coders made an assumption that the target item had been used by the participants. In order to compare the performance of treatment groups on the outcome measures against that of the control or comparison group, as well as group change between pretests and posttests, the meta-analysts used Cohen's  $d$  (Cohen, 1977; Norris & Ortega, 2000). None of the included primary studies actually reported this effect size measure. Therefore,  $d$  was calculated from the reported means and standard deviations and  $t$  or  $F$  values. In one instance, the researchers had to calculate the descriptive statistics themselves from the participants' individual raw scores. For one included study that reported proportions (i.e., the percentage of group members who experienced gain), the meta-analysts adopted an arcsine transformation procedure from Lipsey and Wilson (2001, p. 188). The arcsine value for the corresponding proportion was obtained from the table of arcsine values provided by Lipsey and Wilson (2001, p. 204). In addition to effect sizes, the researchers calculated and reported 95% confidence intervals.

Norris and Ortega (2000) pointed out that the ideal primary research study design for a meta-analysis contrasts a single experimental condition with a single control condition on one dependent variable. Studies with such a simple and straightforward design are rare in the task-based interaction research domain. Most studies included in Keck et al.'s (2006) meta-analysis did not use a true control group but rather included one or more comparison groups that received a variety of treatments. Some of the primary studies did not include the pretest scores needed to calculate gains in scores from the pre- to posttests. In the absence of the true control or comparison group, Keck et al. chose one group as the baseline group so that comparisons could be made between the treatment group(s) and the *baseline* group. It appears that, in some studies, the group that was assigned the status of the baseline group also received a task-based-interaction treatment. The reason it was given the baseline status by Keck et al. was, for example, that the participants received a treatment deemed to be “the least interactive” among all the treatments used in the study or “less than ideal” (e.g., learners were not provided with posttask feedback on the use of the target item). The decision to use one of the interaction groups as the baseline group may have been inevitable. Nevertheless, the fact that some task-based interaction groups were assigned baseline status appears to detract from the purpose of the study that was to compare the effects of task-based-interaction treatments with the effects of treatments not containing such an interaction.

The average effect size computed across all treatment groups was large ( $d = .92$ ); however, there was a substantial variation across treatments in terms of the magnitude of the effects ( $SD = .68$ ). The effects increased slightly over time:  $d = 1.12$  for short-delayed posttests (i.e., 8 to 29 days) and  $d = 1.18$  for long-delayed posttests (i.e., 30 to 60 days).

For the small subset of studies ( $n = 5$ ) that reported both pretest and posttest scores, effect sizes were also calculated for gains as demonstrated on the immediate posttest:  $d = 1.17$  for treatment groups, and  $d = .66$  for control, comparison, or baseline groups, even though the 95% confidence intervals overlapped.

The meta-analysts provided a discussion of the results for each of the coded substantive features of the included primary studies. The calculated effect sizes for different types of target language features were similar:  $d = .94$  for acquisition of grammatical items and  $d = .90$  for acquisition of lexical items. It was not possible to calculate and compare the average effect sizes for specific grammatical or lexical items (e.g., English past tense vs. English reflexive pronouns) because studies investigated a wide range of linguistic features with little accumulation for any given one.

Mean effect sizes for different types of tasks ranged from  $d = 1.6$  (narrative task) to  $d = .78$  (jigsaw task). Contrary to intuitive expectations, tasks in which the target feature was determined by the researchers to be task-essential produced a smaller effect ( $d = .83$ ) than tasks in which the target feature was only task-useful ( $d = .98$ ). Nevertheless, on short-delayed posttests, the mean effect size for task-essential designs was significantly larger ( $d = 1.66$ ) than for task-useful designs ( $d = .76$ ). Tasks involving pushed output (i.e., necessary oral production by the learners) produced larger effects ( $d = 1.05$ ) than tasks without pushed output ( $d = .61$ ) on immediate posttests. The meta-analysts warned that some of these results should be interpreted with extreme caution because, in some instances, the 95% confidence intervals overlapped, and the number of studies with a particular substantive feature was frequently small. Keck et al. expressed confidence that, within the domain included in this meta-analysis, their meta-analytic

results showed that task-based interaction is more effective in promoting acquisition than tasks with little or no interaction.

Keck et al. (2006) summarized current research and reporting practices in the field of task-based interaction and pointed out the following shortcomings: (a) none of the study reports included any measure of reliability for the outcome measures, (b) only 57% of the primary studies reported information about the pretest, (c) two of the studies failed to report the length of the treatment, (d) 62% of the studies failed to set an a priori acceptable probability level, and so on. Most importantly, none of the meta-analyzed study reports provided confidence intervals, standard error of the mean, or effect sizes. Keck et al. also reported that the tests used as the outcome measure varied considerably. Consistent with current research practices in the field, the primary researchers used pretests and posttests that required the participants to make a metalinguistic judgment (e.g., to state whether a certain utterance was grammatically correct), select the appropriate response from several options, or provide a constrained- or a free-constructed response. No reliability was reported for any dependent measures, even though some researchers made references to previous research that cited similar measures as support for the use of these testing measures in their studies.

Based on the analysis of the data obtained through the research synthesis and the quantitative meta-analysis, Keck et al. (2006) provided the following guidelines for future research:

1. Research domain needs to be expanded to include educational settings, learner populations, and TLs that were underrepresented in Keck et al.'s (2006) meta-analysis.

For example, in terms of TLs, the meta-analyzed studies involved only English ( $n = 7$ ), Spanish ( $n = 4$ ), and Japanese ( $n = 3$ ).

2. A greater emphasis needs to be placed on investigations of the effects of learner-to-learner interaction. In the majority of the meta-analyzed studies, interaction treatments involved NS interlocutors who had been trained to carry out specific classroom tasks. The learner participants interacted with other learner participants in only 3 of the 14 included studies (Keck et al., 2006).

3. Research design and reporting practices need to be improved in primary research in the field. Keck et al. (2006) recommended that primary researchers include true control and comparison groups, report descriptive statistics, and compute effect-size measures.

4. More detailed accounts of the interaction that actually takes place during task completion need to be included in primary research reports. Keck et al. (2006) reported that they had to make an assumption that the interaction in tasks had occurred as intended by the task design. Actual conversational exchanges in the classroom may be very different from what the task designers intended (Van den Branden, 2007). Only two of the 14 primary studies included in Keck et al.'s investigation provided analyses of classroom interaction transcripts. Only three of the 14 studies provided counts of the target-item use in the learners' output. If provided, descriptive information of this kind may enable researchers to conduct investigations into what kind of interaction did or did not occur during task completion and for what reason. Such investigations help both task designers and classroom teachers ensure that task-based interaction promotes acquisition of specific TL target features to the greatest degree possible.

Keck et al. (2006) also discussed the need to investigate ways in which interaction effects vary across specific linguistic features (e.g., the past tense “-ed” ending vs. reflexive pronouns in English). As discussed in the subsection titled *Types of Target Structure as Moderator Variables* in this chapter, it is reasonable to assume that task-based interaction affects acquisition of different grammatical structures differentially. The effects for acquisition of individual target structures could not be analyzed by Keck et al. because included primary research studies focused on acquisition of different items, and no systematic comparisons could be made by the meta-analysts. Additionally, many primary study reports offered very few details about the target items.

Unlike Keck et al.’s (2006) meta-analysis, the present meta-analytic study investigated the effectiveness of task-based interaction in acquisition of grammatical TL items only (not lexical items). The mechanisms for grammar acquisition are believed to be different from those involved in the acquisition of lexis and, as reported by Mackey and Goo (2007), effects of interaction on acquisition of grammatical items may be smaller but, once acquisition occurs, these effects may be more durable.

In line with Keck et al.’s (2006) recommendations, the meta-analyst expanded the domain for the present research study to allow for aggregation of larger numbers of studies with similar substantive features. First, studies reported between 2003 and December, 2009 were included. Second, the search procedure included unpublished reports (e.g., doctoral dissertations, master theses, conference reports, etc.) and reports published in professional journals that were not searched by Keck et al. (e.g., *Applied Language Learning*, *Foreign Language Annals*, *French Review*, *Hispania*, *Journal of the Chinese Language Teachers Association*, etc.). These measures yielded an additional

TL that did not appear in studies included in Keck et al.'s meta-analysis (Korean) and some new (i.e., not included in Keck et al.) studies involving learner-to-learner interaction as opposed to NS-led interaction. Limitations of Keck et al.'s meta-analysis are provided in more detail in chapter I of this study. Chapter III outlines the research methodology for the present meta-analysis.

Norris and Ortega's (2000) meta-analysis of effectiveness of L2 instruction is not reviewed here in detail because it focuses broadly on the comparison of the effectiveness of all explicit versus all implicit and all FoF versus all FoFS instructional techniques. A brief review and discussion of Norris and Ortega's meta-analysis from the point of view of the purpose of the present study is provided in chapter I.

Mackey and Goo's (2007) meta-analysis of the effectiveness of conversational interaction in SLA is not reviewed here in detail because it did not focus on the role of focused communication tasks but rather focused exclusively on the learners' opportunities to produce modified output as a reaction to the interactional feedback they received in the process of any classroom interaction (i.e., not specifically task-based interaction). In Mackey and Goo's own words, the researchers focused on the learners' "third turn" in the three-part interactional process of initiation-response-reaction rather than on the presence of the opportunity for the "initial turn" (p. 414). A brief review and discussion of the limitations of Mackey and Goo's meta-analysis relative to the purpose of the present meta-analysis is provided in chapter I.

### Summary

Rich, authentic comprehensible TL input, learner-produced TL output, and possibilities for interaction in the TL with NSs or NNS peers are necessary factors in

successful TL acquisition and the development of communicative competence. Contrary to misconceptions that are sometimes held by FL and L2 teachers, CLT does not reject attention to language form (i.e., teaching of grammar). The concept of communicative competence includes so-called grammatical competence, that is, the degree of grammatical accuracy necessary for successful comprehension and communication of meaning in the TL. Among the three common approaches to FL and L2 instruction (i.e., FoFS, FoM, and FoF), FoF represents the methodological basis for TBLT and is most conducive to developing the learners' communicative competence without sacrificing grammatical accuracy. FoF encompasses a variety of instructional techniques that give appropriate attention to language form while the learners' primary focus is on meaning. One such instructional technique that by design targets the development of mastery of specific grammatical items is so-called focused (structure-based) communication tasks. From the skill-acquisition perspective, focused tasks develop the learners' ability to use target grammatical items at a greater rate and with greater ease while their primary attention is on meaning and not on form, similarly to what happens in real-world TL interaction.

Numerous factors influence the effectiveness of oral interaction that occurs in such focused communication tasks (i.e., task-based interaction). These factors include various task-design features (i.e., type of task, degree of task-essentialness of the target grammatical item, etc.) as well as a wide range of other task-, learner-, teacher-, target-structure-related, and contextual variables. The tests that typically are used in the SLA field to measure acquisition of specific grammatical items involve metalinguistic judgment, selected response, free- or constrained-constructed response, and, occasionally,



oral communication tasks. There are a number of issues surrounding the use of these outcome measures for assessing acquisition of target items, including the reliability and validity of these measures.

The only published meta-analysis that specifically investigated the effectiveness of task-based interaction in acquisition of TL items is Keck et al.'s (2006) meta-analysis. Keck et al. investigated acquisition of both grammatical and lexical items based on a limited number of primary studies published in the top tier of the professional literature. Chapter III outlines the research methodology employed in the present meta-analysis that expanded the domain for such an investigation and focused on a greater number of substantive features and potential moderator variables.

## CHAPTER III

### METHODOLOGY

The purpose of this study was to evaluate the effectiveness of task-based interaction in acquisition of specific grammatical items by meta-analyzing quasi-experimental and experimental studies where treatment involves learner face-to-face interaction in the target language (TL; see Appendix A for a list of abbreviations used in the study) in communicative classroom tasks. In this chapter, the methodology of the study is described including the research design, general characteristics of studies that were included in the study, and procedures that were followed in data collection, pretesting of the coding instrument, and data analysis.

#### Research Design

This study employed the methodology of meta-analysis to summarize and compare the results of quasi-experimental and experimental studies investigating the effectiveness of face-to-face oral interaction in small groups that occurs during TL task completion in classroom foreign language (FL) and second language (L2) instruction of adult learners. In contrast to review of literature, meta-analysis is a type of research synthesis that provides opportunities for a more precise analysis and comparison of primary research study outcomes. According to Lipsey and Wilson (2001), meta-analysis is a form of research that surveys study reports, rather than people. Its major strength lies in the fact that it provides a replicable, statistically grounded summary of research findings by comparing data reported in different primary studies according to a common scale (Norris & Ortega, 2000). This purpose is accomplished by comparing the effect

sizes corresponding to the magnitude of observed relationships reported in the primary studies.

Comparison of the effect sizes across all eligible experimental research in the field allows the researcher to investigate broader research questions that individual studies are not able to address by systematically coding and categorizing features that are common to treatments used in a number of studies and evaluating the impact of these features on learning outcomes (Chaudron, 2006; R. Ellis, 2006; Keck, Iberri-Shea, Tracy-Ventura, & Wa-Mbaleka, 2006; Norris & Ortega, 2006b). Similarly, the potential impact of various research design features on reported outcomes can be evaluated across the research domain as well. Therefore, meta-analysis can help represent research findings in a more differentiated manner than qualitative summaries and so-called vote-counting of statistical significant results reported in primary studies (Lipsey & Wilson, 2001). Qualitative summaries easily are influenced by the reviewer's subjective point of view, whereas reports of statistical significance, albeit more objective, only indicate the degree of probabilistic rareness of a particular outcome, rather than its magnitude and importance (Norris & Ortega, 2000). In addition to overcoming these limitations, meta-analysis also allows the researcher to analyze together the results of individual studies with sample sizes that are too small to render statistically significant findings on their own (Lipsey & Wilson).

The effects of interaction on L2 development have been investigated in several previous meta-analyses conducted in the field of second language acquisition (SLA; Keck et al., 2006; Mackey & Goo, 2007; Norris & Ortega, 2000; Russell & Spada, 2006). None of these studies focused exclusively on acquisition of grammatical TL features

during completion of face-to-face small-group activities that meet stringent criteria for communication tasks presented in chapter II of the present study. The research focus in the meta-analysis completed by Keck et al. was close to the focus of the present meta-analytic study as explained in chapter II; however, like many other meta-analysts in the field, Keck et al. did not include unpublished primary studies. According to Cooper (2003), “it is now accepted practice that rigorous research syntheses include both published and unpublished research” (p. 6). Cooper further explained that synthesists, who are, for example, submitting manuscripts to *Psychological Bulletin*, a premiere publication in the field of psychology, and making summary claims about a particular relationship or treatment, are expected to complete a thorough search of both published and unpublished research.

Due to other limitations of the previously completed meta-analyses (Keck et al., 2006; Mackey & Goo, 2007; Norris & Ortega, 2000) presented in chapter I and to additional opportunities that may have opened up since the time they were conducted, the meta-analyst considered it beneficial to implement another systematic examination of eligible studies. It was expected that inclusion of unpublished study reports, for example, dissertations and conference reports, as well as more recently published studies that generally adhere to more stringent reporting standards may allow for a more detailed analysis of certain moderator variables.

Norris and Ortega (2000) pointed out that the ideal primary research study design for a meta-analysis contrasts a single experimental condition with a single control condition on one dependent variable. Studies with such a straightforward design are rare in the field of language teaching and learning (Norris & Ortega, 2006b). The current

approach to FL and L2 teaching is characterized by so-called principled eclecticism, that is, deliberate and systematic use of a wide range of teaching methodologies and techniques (Kumaravadivelu, 2003). Therefore, according to Norris and Ortega, it is common for primary researchers in the field to pursue similar questions from different methodological perspectives and to incorporate multiple investigative approaches when seeking answers to a complex question. These considerations create additional challenges for the meta-analyst in determining which of the treatments described in the study report fits the definition of task-based interaction and in creating a coding system that will account for various substantive differences between the instructional treatments.

There are other issues specific to the field of language teaching and learning that cause meta-analysts to deviate from the classic guidelines for conducting a meta-analysis as formulated, for example, in Cooper (1998, 2003) and Lipsey and Wilson (2001). One of these issues is that, according to Lazaraton (2000), the popularity of quantitative research in language teaching and SLA has outgrown adequate training in quantitative research methodologies. For this reason, most studies included in Keck et al.'s (2006) meta-analysis did not use a true control group but rather included one or more comparison groups that received a variety of treatments. Studies included in the present meta-analysis were not uniform in terms of the treatment received by the comparison groups either. For example, Toth (2008) compared the effects of learner-led task-based interaction with the effects of teacher-led interaction in the same tasks. Gass and Alvarez-Torres (2005) compared the effects of task-based interaction with the effects of purely input-based activities targeting the same structure (with no interaction) as well as with the effects of input-based activities followed by interaction and vice versa (i.e., input-based

activities preceded by interaction). Notwithstanding such differences, Norris and Ortega (2000, 2006b) believed that a meta-analytic approach can be used in SLA as a systematic means for gathering and analyzing evidence for the purposes of investigating the effectiveness of FL and L2 instruction. The manner in which interpretations of findings based on this evidence can be presented, however, depends upon the degree of adherence to the standards for conducting quantitative research (Norris & Ortega, 2006a).

In the present meta-analysis, the effect sizes for all task-based interaction treatment groups were pulled separately for studies utilizing the treatment and control or comparison group design (i.e., standardized-mean-difference effect sizes) and for the two studies reporting gain scores for a single group of learners (i.e., standardized-mean-gain effect sizes; Adams, 2007; Revesz & Han, 2007). In addition, similarly to Mackey and Goo (2007), within-group gain effect sizes were calculated for all included studies so that the weighted mean for within-group gains would be based on more than two studies. Based on the practice established for meta-analyses in the field of language teaching and learning (Keck et al., 2006; Mackey & Goo, 2007; Norris & Ortega, 2000, 2006b), the overall mean effect-size values for the two subsets of studies were interpreted as suggestive (rather than definitive) findings.

Additionally, it was necessary to analyze various other subsets of individual studies in order to gain insight into the effects of possible moderator variables. Essentially, multiple separate meta-analyses were completed for studies that shared certain coded substantive or methodological characteristics. Analogs to analysis of variance (ANOVA) were conducted to determine whether the variance in effect sizes can be explained by specific moderator variables.

## Data Sources and Search Strategies

The search for the potential candidate studies to be included in the meta-analysis was conducted following these steps:

1. Key and subject word searches within Educational Resource Information Center (ERIC), Linguistic and Language Behavior Abstracts, PsychInfo, Google Scholar, and Dissertation Abstracts for review articles and empirical studies on task-based classroom interaction since 1980. Search terms included combinations of the following terms: *form-focus(s)ed*, *(planned or preemptive) focus on form*, *grammar instruction (or teaching)*, *instructed grammar*, *instructed second language acquisition (SLA; or learning)*, *interaction*, *(communicative or interactive) tasks*, *task-based*, *focus(s)ed (communication or communicative) tasks*, *structure-based (production, communication, or interactive) tasks*, *grammaring tasks*, *(collaborative or structured) output*, *(focus-on-form) output processing tasks*, *effects (or effectiveness)*, *empirical (quasi-experimental or experimental)*, *results*, *(literature) review*, *target structure*, and so forth. Some of the terms, such as *morphosyntactic development* were added later after the search procedure had been started based on identified recurring key words in potentially eligible research studies. When the search yielded an unpublished report (e.g., doctoral dissertation) that was not available on the web, an electronic mail request was sent to the author (see Appendix D).

2. Manual search of the following journals in the field: *Applied Language Learning*, *Applied Linguistics*, *Applied Psycholinguistics*, *Asian English as a Foreign Language (EFL) Journal*, *Association Internationale de Linguistique Appliqué (AILA) Review*, *Canadian Modern Language Review*, *Die Unterrichtspraxis - Teaching German*,

*English Language Teaching (ELT) Journal, French Review, Foreign Language Annals, Hispania, Japanese Association of Language Teachers (JALT) Journal, International Journal of Educational Research (IJER), International Review of Applied Linguistics (IRAL), Journal of the Chinese Language Teachers Association (JCLTA), Language Awareness, Language Learning, Language Teaching, Language Teaching Research (LTR), Modern Language Journal (MLJ), Regional Language Center (RELC) Journal, Second Language Research (SLR), Studies in Second Language Acquisition (SSLA), System, and Teachers of English to Speakers of Other Languages (TESOL) Quarterly.*

Many of these publications were not included in the manual and electronic journal search in Norris and Ortega's (2000), Keck et al.'s (2006), or Mackey and Goo's (2007) meta-analyses.

3. Examination of bibliography sections of textbooks and seminal volumes in the SLA field (Doughty & Williams, 1998; R. Ellis, 1994b, 2001, 2003; Gass & Mackey, 2007; Gass & Selinker, 2008; Larsen-Freeman, 2003; Larsen-Freeman & Long, 1991; Leaver & Willis, 2004; Mackey, 2007; Mitchell & Myles, 1998; Nunan, 1999) as well as milestone review articles (N. Ellis, 1995; R. Ellis, 2006a; 2006c; Keck et al., 2006; Nassaji, 1999; Norris & Ortega, 2000; Pica, Kang, & Sauro, 2006; Spada, 1997) for references to relevant study candidates.

4. Once the review articles and empirical studies were identified, the references sections of these sources were searched for additional studies.

5. The list of identified studies was submitted for review to two experts in the field in an attempt to identify possible omissions. The experts' responses indicated that they did not identify primary studies that should be added or removed from the list.



### Fail-Safe N

Because there is a marked tendency among researchers to report and publish statistically significant results, there is a potential for the reporting and publication bias, that is, nonstatistically significant results are much less likely to be retrieved than statistically significant results (Cooper, 1998; 2003). The so-called *file-drawer effect* was partially addressed in the present meta-analysis by inclusion of unpublished studies. Nevertheless, outside of research study reports completed as dissertations, researchers may still choose not to make public the results that did not reveal a statistically significant relationship between variables. Because the number of the studies identified as meeting the criteria for inclusion and exclusion was small for this meta-analysis, there potentially is a threat to the external validity and generalizability of the findings with regard to the relationship between task-based interaction as instructional treatment for teaching target grammatical structures and learners' acquisitions of these structures.

Therefore, after statistically significant overall effect sizes were found as reported in chapter IV, it was important to calculate the so-called *fail-safe N*. This calculated number shows how many unretrieved studies supporting the null hypothesis would have to be located to counteract the conclusion that a statistically significant relationship exists (Lipsey & Wilson, 2001; Rosenthal, 1991).

The results of the *fail-safe N* statistic indicate that approximately 328 studies with a null result would be required to reduce the effect size to a nonsignificant level for standardized-mean-difference effect size. Considering the fact that intensive electronic and manual searches of the literature produced only 15 studies that met the inclusion and exclusion criteria for this meta-analysis, it seems highly unlikely that such a great number

of studies existing in researchers' file drawers do not appear in the databases of published and unpublished sources and were not located by manual searches.

### Inclusion and Exclusion Criteria

The following inclusion and exclusion criteria were applied to the 386 studies identified through the search using the keywords. Fifteen of these studies met the inclusion and exclusion criteria.

#### *Inclusion Criteria*

A study was included in the meta-analysis if it met all of the criteria listed below.

1. The study investigated acquisition of specific FL or L2 target grammatical structures.
2. The independent variable was the treatment and had the use of oral communicative form-focused tasks as one of the levels.
3. The dependent variable was the learners' scores on a posttest that aimed to measure the acquisition of the grammatical structure(s) targeted by the treatment.
4. The tasks used in the treatment were designed specifically to foster meaningful interaction in pairs or small groups involving the use of the specified grammatical features.
5. The tasks used in the treatment met the criteria for tasks that have been delineated in chapter II of the present study.
6. The report contained an adequate description of the tasks employed as treatment so it was possible to ascertain that these activities do indeed meet the criteria for tasks.

7. The study investigated language acquisition in adults or postpubertal adolescents.

8. The study was published or reported between 1980 and December 2009. The year 1980 is a common starting point because in that year Long (1981, 1996) formulated the interaction hypothesis that posits, among other things, that interaction is crucial to TL acquisition (Keck et al., 2006). Long (1982) also formulated the tenets of Focus on Form (FoF) as a key methodological principle of task-based language teaching (TBLT).

### *Exclusion Criteria*

All potential candidate studies also were checked against the exclusion criteria listed below. Studies were excluded if

1. Studies utilized descriptive or correlational designs.
2. Studies involved online or computer-assisted language learning.
3. Studies involved only written tasks (however, learners could be taking notes, preparing lists, writing down arguments, etc. as part of completing oral tasks).
4. Treatment used in the study only contained so-called consciousness-raising tasks, hypothesis-building, or other metalinguistic problem-solving activities (e.g., dictogloss, text-reconstruction activities, etc.) during which learners talked *about* grammar either in their first language (L1) or TL, rather than exchanged real-world (vs. metalinguistic) information in the TL.
5. Studies utilized form-focused communication tasks for a different purpose, for example, for linguistic analysis of learner discourse generated during these tasks or for the sole purpose of investigating effectiveness of corrective feedback.

Based on these inclusion and exclusion criteria, only three of the 13 primary studies included in Keck et al.'s (2006) previous meta-analysis of effectiveness of task-based interaction, specifically, Iwashita (2003), Loschky (1994), and Mackey (1999), qualified for inclusion in the present study. Only four studies out of the 27 studies in Mackey and Goo's (2007) meta-analysis were included. These were the same three studies as in Keck et al. as well as a study by Gass and Alvarez-Torres (2005) that was published after Keck et al.'s meta-analysis was completed.

### Coding

The instrument was a researcher-designed coding form (see Appendix C). All identified studies that meet both the inclusion and the exclusion criteria were coded by the researcher and an individual experienced in the interpretation of primary research studies in the field of FL and L2 acquisition.

The coding form was tested by the researcher on a few studies prior to the beginning of the coding process to learn if it needs to be modified in any way. Some modifications to the coding form were made in the process of coding as a result of discussions with the second coder. For example, a section for learner characteristics of the entire sample (i.e., all the groups involved in the study) was added. (Originally, the coding form only contained a learner-characteristic section for each group; however, these data were not necessarily provided in the primary studies for each group separately.) Additionally, the coding form was modified when unanticipated levels of the variables were encountered during coding, for example, *morphosyntactic* structures (i.e., language structures that combine morphological and syntactic features). The coding categories such as study identification information, characteristics of the outcome

measure, methodological features, learner characteristics, treatment design, and pedagogical features as well as features related to the quality of study are outlined in the subsequent subsections.

### *Study Identification Information*

In this category, the background characteristics of the research study were coded. These characteristics included the author's name, the title of the publication or report, the publication or reporting year, and the source of the study (e.g., the publication or database from which it was obtained).

### *Characteristics of the Outcome Measure*

The studies were coded based on the dependent variable (i.e., scores on the immediate and delayed posttests or gain scores). In case more than one category of the outcome measure was used in the primary study, all relevant information about the types of tests and the dependent variables were noted. Meta-analytic findings were reported separately for the subset of studies that investigated acquisition of the target grammatical structures based on the differences between the control (or comparison) and treatment groups and for within-group gains (i.e., gains in learner scores from the pre- to the posttests) across all the studies (see subsection titled *Effect-Size Measures*).

Nevertheless, there was an issue of different types of pre- and posttests being used in both of these two subsets of primary research studies. Cooper (2003) warned against including primary studies that use different types of posttests in one meta-analysis. In the field of language teaching and learning, however, it is common for teachers and researchers to design unique pre- and posttests, especially when acquisition of a specific language item (in a specific TL) is being investigated (Chaudron, 2003; Erlam, 2003;

Gass & Mackey, 2007; Larsen-Freeman & Long, 1991; Norris & Ortega, 2000, 2006a). Therefore, included primary research studies investigating the effects of task-based interaction on acquisition of specific grammatical structures used such unique tests designed specifically for these studies ranging from “discrete-point recognition items to full-blown spontaneous communicative performance” (Norris & Ortega, 2006a, p. 729).

Established standardized tests are rare in the field of FL and L2 teaching and learning, and they are proficiency-oriented in nature (e.g., Test of English as a Foreign Language [TOEFL]), which makes them unsuitable for measuring mastery of specific language items (Gass & Mackey, 2007; Gass & Selinker, 2008). In designing tests that measure acquisition of targeted language items, primary researchers frequently go to great lengths to ensure that the TL material used in such tests does not present any additional challenges to the learners by verifying, for example, that all the vocabulary and nontargeted grammatical structures involved are already known to the learners (Gass & Mackey, 2007; Gass & Selinker, 2008). This practice helps reduce construct-irrelevant variance in learner scores; however, it also results in the creation of tests that are suitable for a particular group of learners who are using a particular curriculum and, unlike proficiency language tests, cannot be used for other groups of learners.

Additionally, some primary research studies utilize so-called custom-made tests that are designed for individual learners by the researchers based on the errors made by these learners on the pretest conducted before the instructional treatment or during the task-based-interaction treatment (Adams, 2007; Egi, 2007; Kowal & Swain 1994; Swain & Lapkin 1998, 2001, 2002). Although the use of custom-made tests makes sense from the point of view of understanding the deeply individualized nature of learners’

interlanguage development, it raises serious issues from the point of view of measurement when comparisons across studies are made (Mackey & Goo, 2007). Only one of the studies employing custom-made posttests met the criteria and was included in the present meta-analysis (i.e., Adams, 2007).

Cooper (2003) suggested that, in order to overcome the issue of different types of outcome measures used (i.e., posttests), several separate meta-analyses be completed within the same research synthesis in order for the meta-analyst to be able to make summary statements about relationships between the variables. Following the established practice in the field of language teaching and learning (Keck et al., 2006; Mackey & Goo, 2007; Norris & Ortega, 2000, 2006b), the overall mean effect size for the task-based interaction obtained from all included primary studies (regardless of the type of posttest used to measure acquisition of the target structure) was reported but interpreted as a suggestive, rather than a definitive, finding. The differences between specific types of outcome measures were treated as potential moderator variables.

The characteristics of the outcome measures (i.e., immediate and delayed posttests measuring acquisition of the target structures) as well as of the pretests were coded using the following categories presented in chapter II under *Types of Outcome Measures as Moderator Variables*: (a) metalinguistic grammaticality judgment, (b) selected response, (c) constrained-constructed response, (d) free-constructed response, and (e) oral-communication task. For the most part, this classification is based on the categories used in Norris and Ortega's (2000) meta-analysis, with the exception of the oral-communication task category. This latter category was used if the participants were required to engage in interaction with each other or the teacher, researcher, or tutor, for a

nonlinguistic real-world purpose during the test and this activity met the criteria for a task that are presented in chapter II.

As pointed out in chapter III in the *Characteristics of the Outcome Measure*, there was a great deal of variation in the length of time elapsed before the administration of the delayed posttest in the included studies ( $M = 33.42$ ; range from 7 to 120 days). Posttests with a delay of 0 to 27 days ( $k = 6$ ) were considered short-delay posttests, and those with a delay of 28 to 120 days ( $k = 6$ ) were considered to be long-delay posttests in the present meta-analysis.

Additionally, it was recorded whether the outcome measure(s) was or were congruent with the task-based teaching methodology, for example, a metalinguistic-judgment or selected-response measure is not congruent with task-based instructional treatment that the examinees have received whereas an oral-communication task or free-constructed response are congruent with it. Presence or absence of counterbalancing measures in the pre- and posttests (i.e., whether there was an attempt to control for the so-called test learning effects, etc.) also was coded. Counterbalancing measures included not using the same tasks in the pre- and posttest as well as controlling for the task order effect by not presenting test tasks in the same order to all examinees (Mackey & Gass, 2005).

### *Methodological Features*

In this category, type of publication or source, type of outcome measure (e.g., standardized test, uniform researcher-made test, or custom-designed researcher-made test), treatment duration, educational setting (i.e., high school, university, adult education, Intensive English Program [IEP], English for Specific Purposes [ESP] program, etc.) were recorded. Research design features such as participant selection and assignment



(i.e., random, intact class, or volunteer), presence or absence of the control or comparison group(s), the number of participants in the treatment, control, and comparison groups as well as the basis for determining participant TL proficiency levels (i.e., impressionistic judgment, institutional placement test, standardized test, etc.) also were reported.

Studies were coded for the presence or absence of a pretest, whether any individuals were eliminated on the basis of the pretest, and for what reason (e.g., those who already demonstrated mastery of the target structure). Such features as the presence of an immediate posttest, a delayed posttest, and how much time expired before the delayed posttest were coded as well.

Information about the TL was reported in the following manner: (a) name of the language (e.g., English, Spanish, Japanese, etc.), (b) whether it was being studied as an L2 or FL (i.e., within vs. outside of the target culture environment), (c) for languages other than English, their group number based on MacWhinney's (1995) classification (e.g., Group I for Spanish, Group II for German, Group III for Hungarian, Group IV for Japanese, Group V for Georgian, etc.). The degree of dissimilarity between the TL and the learners' L1 was noted as well, wherever possible, based on their relative closeness to each other using a similar system. For example, for English-speaking learners of Hungarian or for Hungarian-speaking learners of English, the degree of dissimilarity between the two languages was marked as III, and for English-speaking learners of Japanese or Japanese-speaking learners of English, it was marked as IV.

The statistical outcomes of the primary studies were reported: (a) the means, standard deviations, and sample sizes for each group in a comparison or hypothesis test;

(b)  $t$  tests or  $F$  tests and the associated degrees of freedom; (c) and exact  $p$  level and sample size; and (d) proportions of learners who experienced gains.

### *Learner Characteristics*

This category involves the characteristics of the participants in the primary studies. These characteristics included the participants' average age, gender, L1, and TL proficiency level (i.e., beginning, low intermediate, high intermediate, advanced) for both treatment and control or comparison groups.

### *Treatment Design and Pedagogical Features*

This category provided information about the task(s) used as instructional treatment such as (a) task description (spotting the differences between pictures, negotiating a joint decision, etc.), (b) type of task (information-gap vs. opinion gap, divergent vs. convergent, etc.), (c) origin of task (i.e., whether the treatment tasks were designed by the researcher, the classroom teacher, or a curriculum development specialist); and (d) whether the task(s) were administered by the regular classroom teacher, the researcher, an assistant who is a native speaker (NS) of the TL, and so on. Any information that was available regarding the presence of teacher or teaching assistant (TA) training in the use of treatment tasks as well as teacher or learner beliefs and attitudes (i.e., presence of information about teacher familiarity with task-based language teaching in general and perceptions regarding its effectiveness in teaching grammar in particular) was recorded as well when available.

Information regarding the target grammatical structure was coded as follows: (a) name of the structure, (b) type of the structure (morphological vs. syntactic vs. morphosyntactic), and (c) any other information that was provided or could be inferred

from the study (simple vs. complex, ambiguous vs. unambiguous, etc.). Also recorded was the degree of task-essentialness of the target structure (i.e., task-essentialness, task-usefulness, or task-naturalness) and whether evidence of target structure use (e.g., through interaction transcripts, usage counts, etc.) was available. Additionally, any information regarding the presence of priming for target structure use during the pretask phase (e.g., rule review, modeling, etc.), presence of monitoring for accuracy of use of the target structure and error correction during task completion, and presence of target structure focus during the posttask phase (e.g., rule review, error treatment) was noted as well.

### *Quality of Study*

Quality of study codings reflected peer review process (e.g., peer reviewed or not peer-reviewed) and attrition rates for control, comparison, and treatment groups. Information on reliability and validity of testing instruments used as outcome measures reported in the primary research studies or the absence thereof were recorded as well.

Rosenthal (1991) suggested using a weighting system that takes into account the methodological quality of the studies included in a meta-analysis. Weighting was not used in the present meta-analysis due to the great diversity of primary study designs. Additionally, methodological quality of empirical research in SLA generally is considered to be lower than in the field of cognitive psychology; therefore, it would be challenging to devise an effective weighting scheme.

### *Validity and Reliability of the Meta-Analysis*

In addition to the validity and reliability issues that are present in primary research that influence meta-analytic findings, meta-analysis as a type of research is

prone to its own threats to reliability and validity (Cooper, 1998). For example, it is important that meta-analysts take steps in order to diminish the potential effects of so-called expectancy bias that can lead to subjective interpretations of findings (Cooper, 1998; Norris & Ortega, 2000). Some of these steps relevant to the present study are described in the subsequent sections.

### *Validity*

Validity is defined by Cooper (1998) as trustworthiness of the many decisions made at every stage of the meta-analysis. Considerations related to the validity of the present meta-analytic study are provided in this section.

In line with Cooper's (1998) and Norris and Ortega's (2000) recommendations, the coding of the primary study reports was conducted after the definitions for various substantive and methodological features had been established as presented in chapter II of this study. Nevertheless, because both the researcher and the additional coder had used focused communication tasks in their teaching and had advocated the use of such tasks while conducting language teacher training (as explained in the section titled *Qualifications of the Researcher*), there was a potential for bias favoring task-based interaction. Therefore, maintaining the objectivity of the coding process to minimize the expectancy bias was a major focus of the training sessions for the second coder and of all discussions between the researcher and the second coder.

From the beginning, it was expected that there was a possibility that some primary research studies will not provide sufficient information about important characteristics of the treatment, learners, the target structure, outcome measures, and so forth. Incomplete reporting by the primary researchers consequently could jeopardize the validity of this

study. When essential statistical information or information about the treatment was missing (e.g., there is not sufficient information that the treatment fits the criteria for a communicative focused task), a reasonable effort was made to contact the primary researcher and obtain the missing information. Ultimately, in situations where crucial information could not be inferred or obtained directly from the author of the study, the study had to be dropped from a particular part of the analysis (i.e., for a specific moderator variable) due to insufficient information. In other instances, the meta-analyst had to exercise caution by using merely suggestive, descriptive statements about the effects of independent and moderator variables.

Similarly, due to the deviations from the classic meta-analytic procedures that are common in the SLA field as explained in sections titled *Research Design* and *Characteristics of Outcome Measures*, any generalizations that were made across different types of learners, outcome measures, specific characteristics of instructional treatments, and so forth were interpreted as merely suggestive, rather than definitive. Additionally, because methodological inconsistencies are widespread within the research domain (Lazaraton, 2000; Norris & Ortega, 2000, 2006a, 2006b), the present meta-analytic study adopted an inclusive approach with regard to studies that generally fit the criteria for inclusion but have certain methodological flaws (e.g., absence of control groups, absence of randomized sampling of study participants, etc.). Plonsky (2010) provided meta-analytic evidence of a relationship between study quality and effect-size outcomes. Therefore, the inclusion of primary studies of lesser quality may limit the generalizability of the findings of the present study. The meta-analyst made every effort

to investigate the effects of various methodological (i.e., research design and quality of study) features on primary research study outcomes in the present study.

### *Interrater Reliability*

The coding process was conducted according to a series of stages that were meant to ensure that checks on the researcher's judgments were included. To ensure reliability of coding, an additional coder was recruited from the ranks of the individuals who are familiar with primary research practices in the field of SLA and, in particular, form-focused instruction (FFI) and TBLT. To maximize the level of interrater reliability, the purpose and the rationale for the study were explained to the additional coder who then was trained on how to code the assigned studies using the coding form.

After coding three studies independently, the two coders compared the completed coding forms, resolved all disagreements through discussions, and refined the coding categories when necessary. The remaining studies were split in half and coded by the two coders independently. Rather than wait for both coders to complete the coding, every third study was duplicated and given to the other coder. This measure allowed for prompt checking for reliability throughout the coding process and helped identify potential sources of ambiguity and difficulty. After all the studies included in this reliability check were coded, the percent agreement ratio was calculated and was equal to 81.08%. (It had been established a priori that an agreement ratio of 80% or higher will be considered acceptable.) All differences in coding between the two coders were discussed in subsequent meetings and resolved by consensus.

### *Pretesting of the Coding Form*

Prior to the start of the coding process, the second coder received a training

session that consisted of review and discussion of the coding categories. A sample primary research study that was not going to be included in the meta-analysis was used to provide hands-on coding practice for the second coder.

After the training session, in order to assess the reliability of the coding process, three randomly selected studies were coded independently by the researcher and the second coder using the original researcher-designed coding form. The two coders then examined the coding data together and discussed any disagreements. The coding procedure was revised somewhat based on the suggestions that emerged during the discussion of the coding protocol and of the categories that were used as coding options for specific study features.

### Data Analysis

In order to address the research questions, effect sizes obtained from the primary study reports were compared. The overall effect size and 95% confidence intervals were computed. When the effect sizes calculated for the target structures in the primary studies were tested for homogeneity using the  $Q$  statistic, they were found to be heterogeneous (see a more detailed discussion of heterogeneity of effect sizes in the next subsection). Therefore, more fine-tuned analyses were performed with effect sizes aggregated according to distinct coded features (e.g., different kinds of treatments, outcome measures, learner populations), and comparisons were drawn between these average effects. The following subsections outline the statistical procedures that were used in the meta-analysis.

### *Effect-Size Measures*

Effect sizes could not be extracted directly from the primary study reports because

none of the studies reported such outcomes. Therefore, effect sizes were calculated based on the data provided in the reports. Cohen's  $d$  (1977) was used in order to compare the performance of treatment groups against the performance of the control and comparison groups on the outcome measures. For the subset of primary studies that reported group change between pretests and posttests, Cohen's  $d$  was used as well to compare the magnitudes of the gains.

The  $d$  index, or the standardized mean difference (Cohen 1977), was calculated by subtracting the mean value of the control or comparison group on the dependent measure from the mean value of the treatment group on the same measure and then dividing the difference by the pooled standard deviation of the two groups. If the primary study investigated pretest to posttest differences within a single group (i.e., repeated measures design), Cohen's  $d$  was calculated on the basis of the mean gain from the pretest to the posttest for a single group on a single measure by dividing the difference between the mean posttest and pretest values by the pooled standard deviation of the pre- and posttest groups. Because this within-group estimate is not comparable with between-group estimates, the within-group and between-group mean gain effect-size values had to be treated separately (Norris & Ortega, 2000).

If a study reported the group means and standard deviations, Cohen's  $d$  was calculated from these reported values. If a study reported proportions (i.e., the percentage of group members who experienced gain), the arcsine transformation procedure suggested by Lipsey and Wilson (2001, p. 188) was used. For example, arcsine transformations were performed to obtain the effect size for Mackey's (1999) study as well as for the oral-production test for English questions in Kim's (2009) study.



Hedges (1981) pointed out that the  $d$ -index tends to be biased upwardly when based on small sample sizes. Therefore, after the  $d$  values were obtained, they were converted to Hedges's  $g$  values, that is, unbiased estimates (Hedges & Olkin, 1985, p. 81).

Any treatments other than task-based-interaction treatments were considered to be comparison treatments for the purposes of this meta-analysis, regardless of the primary researcher's own designation. Lipsey and Wilson (2001) recommended combining within-study effect sizes in order to avoid the problem of nonindependence of effect-size values. For this meta-analytic study, however, if a particular primary study used different treatments, types of tasks, or target structures, combining the within-study effect sizes in all situations would be counterproductive to the purpose of the study. For example, if the effect sizes always were combined for different types of tasks and different types of target structures, it would be impossible to analyze the information about how the characteristics of each task and each targeted structure impacted the effect of the experimental treatment (Norris & Ortega, 2000, 2006b).

Following this consideration, pooled means were calculated across all treatment or all comparison groups to be used in the effect-size calculation in order to address Research Questions 1 and 2 (for studies that employed more than one experimental treatment group or more than one comparison group). To address Research Questions 3, 4, and 5, however, for studies that involved multiple treatments or multiple outcome measures, separate effect sizes were entered for each such variable and used in the analysis of the effects of moderator variables as presented in the *Moderator Variables* section.

Most importantly, in accordance with established practices in the SLA field (Keck et al., 2006; Mackey & Goo, 2007; Norris & Ortega, 2000; 2006; Spada & Tomita, 2010), the effect sizes within studies with multiple target structures (i.e., Adams, 2007; Gass & Alvarez-Torres, 2005; Iwashita, 2003; Jeon, 2004; Nuevo, 2006) typically were not averaged within the study. Effect sizes associated with different target structures were treated as independent effect sizes for the most part in order to avoid obfuscation of the effects of important variables (Norris & Ortega, 2006b).

Nevertheless, in adherence with practices established in the field of cognitive psychology, in some instances, findings also are reported in the present meta-analysis taking into account only a single effect-size value associated with one randomly selected target structure for each of the five included studies that investigates acquisition of multiple target structures. Lipsey and Wilson (2001) suggested two possible options for dealing with a situation where “research studies under review include multiple effect sizes for the same conceptual relationship” (p. 125). The first option is to select one of the effect sizes randomly or on the basis of some criteria. The second option is to average the effect-size values. Because effect sizes associated with different grammatical structures typically are not averaged in the task-based interaction research domain, the former option was selected in the present meta-analysis. Random selection appeared to be reasonable because choosing an effect size associated with a particular target structure based on some criteria would contradict the research purpose of investigating characteristics of the target structure as potential moderator variables. The weighted mean effect-size values calculated on the basis of only one randomly selected target structure per study were reported primarily for informational purposes as well as to demonstrate

that the resulting weighted mean effect sizes were similar to the values obtained when treating all effect sizes associated with different target structures within a study as independent primary units of analysis in accordance with the common practice in the domain.

Effect-size calculations that were completed on the basis of statistical data provided in the 15 included studies yielded effect sizes associated with 22 target structures whose acquisition was investigated in these studies. The overall weighted mean effect-size values that are reported in chapter IV are based on a total of 926 learner participants ( $k = 15$ ), where  $k$  is the number of studies included in the meta-analysis. In accordance with the research purpose, two types of effect-size values were computed: (a) the standardized-mean-difference effect size that is based on the between-groups contrasts (i.e., differences in performance between the experimental and control, or experimental and comparison groups on immediate and delayed posttests) and (b) the standardized-mean-gain effect size that is based on the within-group contrasts (i.e., differences between the learners' performance on the pretest vs. the immediate or delayed posttest). For Research Question 1, separate standardized-mean-difference effect sizes were computed for the contrast between the performance of the experimental and control groups, on the one hand, and the experimental and comparison groups, on the other hand (if a comparison group or groups were present in the study).

Only two of the included studies (Adams, 2007; Revesz & Han, 2006) did not involve a control or comparison group; therefore, only the standardized-mean-gain effect size could be calculated for these studies. Because meaningful comparisons cannot be made on the basis of only two studies, standardized-mean-gain effect sizes also were

calculated for the other studies (in addition to standardized-mean-difference effect sizes), all of which involved a posttest, thus making these calculations possible. Pretest-posttest comparisons, however, tend to produce larger effects (Morris, 2008), and the associated effect sizes tend to be biased, or inflated (Cheung & Chan, 2004; Gleser & Olkin, 2009). Therefore, such within-group comparisons should be treated with caution. In general, the findings provided below are a result of the meta-analysis of only 15 studies in the domain and, therefore, should be interpreted as merely suggestive rather than definitive.

### *Nonhomogeneity of Effect Sizes*

The effect sizes were tested for homogeneity using Hedges's (1981)  $Q$  statistic and were found to be heterogeneous. In general, the purpose of the homogeneity test is to make sure that possible presence of extreme values that may not be representative of the population does not influence the findings of the meta-analysis disproportionately (Lipsey & Wilson, 2001). In the present study, the meta-analyst employed a number of techniques recommended by Lipsey and Wilson in order to attempt to remove outliers (i.e., effect sizes representing extreme values) and thus to find a homogeneous set of effect sizes. The following methods and combinations thereof were attempted: (a) removing values that were considerably greater or considerably lower than the other values, or both, and (b) adjusting considerably greater values to less extreme values. The latter was attempted by means of *Windsorizing*, that is, recoding the extreme values to more moderate ones (Lipsey & Wilson). For example, all effect sizes greater than 2.00 were recoded as 2.00 and all effect sizes greater than 3.00 were recoded as 3.00; however, all attempted variations of this procedure failed to result in a homogeneous set. Additionally, the meta-analyst attempted to remove studies with low numbers of

participants (e.g., fewer than 10 or 12 participants in the control or experimental groups). These attempts largely were unsuccessful because, even if a homogeneous set could be found, it typically could be accomplished only by removing far too many effect-size values from the set, thus further reducing the number of included studies. The *Test of Homogeneity* section in chapter IV presents the meta-analyst's most successful attempts to arrive at homogeneous sets of effect-size values both for the standardized mean difference and the standardized mean gain.

Because achieving homogeneity resulted in drastic reduction of already scarce data, the analysis was conducted based on the original (heterogeneous) sets of effect sizes in the present study. Large variability in effect-size values and presence of extreme values are well-documented in the SLA field (Mackey & Goo, 2007; Norris & Ortega, 2000; Plonsky, 2010) and possibly are the reason for the fact that tests of homogeneity typically are not reported in meta-analyses in the domain. The presence of considerable variability in effect sizes across studies confirmed the need for the analysis of potential moderator variables (Lipsey & Wilson, 2001) as outlined in Research Questions 3, 4, and 5.

A separate average effect size was computed for delayed posttests for the subset of studies that used a delayed posttest. Based on the guidelines suggested by Cohen (1977), the effect size of .20 was interpreted as small, .50 as medium, and .80 as large. In addition to the average effect sizes, the researcher calculated 95% confidence intervals and checked for overlaps.

#### *Moderator Variables*

As evident from the Research Questions, the purpose of the study was not only to

establish the overall relationship between the independent variable (i.e., task-based interaction) and the dependent variable (i.e., acquisition of specific TL grammatical structures; Research Questions 1 and 2) but also to investigate the factors that were associated with variations in the magnitudes of the relationships between these two variables, that is, the so-called moderator variables (Rosenthal, 1991).

In line with Research Question 3, the type of task used as the instructional treatment (e.g., information-gap vs. reasoning gap) was an important moderator variable. Effect sizes obtained from primary studies that involved different task types were aggregated and compared in order to examine the role of this moderator variable (i.e., task type) provided that there was sufficient accumulation for each of the subtypes.

In line with Research Question 4, other coded features were examined as possible moderator variables in a similar manner by aggregating and comparing the corresponding effect sizes. These features included such variables as the type of grammatical structure being targeted by the treatment, duration of instruction, as well as miscellaneous other task-, teacher-, learner-related, and contextual variables.

Finally, in line with Research Question 5, the effect of such moderator variable as the type of outcome measure was examined similarly. In particular, it was investigated whether outcome measures that are congruent with the task-based methodology (i.e., measures that use communicative assessment tasks) resulted in larger effect sizes as compared with measures that use more traditional, noncommunicative, assessment items.

#### Qualifications of the Researcher

The researcher obtained a Bachelor of Arts degree in teaching foreign languages (English and German) from the Moscow State Linguistic University in Moscow, Russian

Federation, in 1986. In 2004, she earned a Master of Arts degree in teaching foreign language (Russian) at the Monterey Institute of International Studies, now an affiliate of Middlebury College, in Monterey, California.

The researcher has extensive experience in teaching FL to adult learners at all proficiency levels (i.e., beginning, intermediate, and advanced). She taught English as FL in Russia between the years of 1986 and 1993 as well as Russian as FL in the US between the years of 1991 and 2004. Additionally, during the period of 2001-2004, she served as department chair (i.e., teacher supervisor) of one of the Russian language departments at the Defense Language Institute Foreign Language Center (DLIFLC) in Monterey, California. During the period of 2004-2007, she worked as a faculty development specialist at the Faculty and Staff Development Division that provided pre- and in-service language teacher training to DLIFLC faculty members representing over 30 languages. In 2007-2009, she served as academic specialist responsible for faculty development at the Middle East School III at DLIFLC working with over 100 teachers of Modern Standard Arabic. Since March 2010, she has been serving as academic specialist for curriculum development at the Student Learning Center at DLIFLC.

The researcher has implemented a variety of focused (structure-based) communication tasks in her teaching of both English and Russian. She also has conducted training in applying TBLT for teachers and program managers from various language programs at DLIFLC. In 2003, she conducted a quasi-experimental research study investigating the effectiveness of various types of FFI for her Master's program course project. She has co-conducted TBLT-related presentations at DLIFLC training events and

national conferences, for example, at the 2007 annual convention of Teachers of English to Speakers of Other Languages (TESOL) and the 2007 International TBLT Conference. Her co-presenter, Natalie Lovick, currently is serving as academic specialist at the European and Latin American School at DLIFLC. The researcher and Natalie Lovick have co-authored articles on TBLT and FFI that appeared in the DLIFLC *Bridges* publication.



## CHAPTER IV

### RESULTS

The purpose of this study was to investigate the effects of learners' task-based interaction in face-to-face focused oral-communication tasks on the acquisition of specific grammatical structures of the target language (TL; see Appendix B for a list of abbreviations used in this study). This meta-analysis synthesized the results of 15 published (journal articles and chapters in edited volumes) and unpublished (dissertations) quasi-experimental studies in the task-based interaction domain. Effect-size calculations that were completed on the basis of statistical data provided in the 15 included studies yielded effect sizes associated with 22 target structures whose acquisition was investigated in these studies. The 15 primary studies included a total of 926 participants who were learning a number of TLs. The majority of the studies included a control group and a pretest as well as an immediate posttest, a delayed posttest, or both. The outcome measures (i.e., measures of acquisition of target structures) used in the primary studies ranged from those requiring the learners to state whether or not the target structure was used correctly in a sentence to oral-communication tasks. In accordance with the research purpose, two types of effect-size values were computed: (a) the standardized-mean-difference effect size and (b) the standardized-mean-gain effect size using the procedures presented in chapter III of the present study. These weighted mean effect-size values were calculated separately for immediate and delayed posttests. Table 5 presents an overview of the included studies including some of their research design characteristics and target grammatical structures whose acquisition they investigated.

Table 5  
Overview of 15 Studies Included in the Present Meta-Analysis

Study/Target Structure	Publication Type	n	Control Group	Pretest	Immediate Posttest	Delayed Posttest
Adams (2007) Questions (English) Past Tense (English) Locatives (English)	Chapter	25	na	na*	yes	na
Gass & Alvarez-Torres (2005) Gender Agreement (Spanish) “Estar” + location (Spanish)	Article	102	yes	yes	yes	na
Horibe (2002) Temporal Subordinate Conjunctions (Japanese)	Dissertation	30	yes	yes	yes	yes
Iwashita (2003) Locative Constructions (Japanese) Verbal “-te” morpheme	Article	55	yes	yes	yes	yes
Jeon (2004) Honorifics (Korean) Relative Object Clauses (Korean)	Dissertation	34	yes	yes	yes	yes
Kim (2009) Past Tense Questions	Dissertation	191	na	yes	yes	yes
Koyanagi (1998) Conditional “to” (Japanese)	Dissertation	31	yes	yes	yes	yes
Loschky (1994) Locative Constructions	Article	41	na	na**	yes	yes
Mackey (1999) Questions (English)	Article	34	yes	yes	na	yes
Nuevo (2006) Past Tense (English) Locatives (English)	Dissertation	103	yes	yes	yes	yes

continued on the next page

Table 5 continued

Study/Target Structure	Publication Type	n	Control	Pretest	Immediate Posttest	Delayed Posttest
Revesz (2007) Past Progressive (English)	Dissertation	90	yes	yes	yes	yes
Revesz & Han (2006) Past Progressive (English)	Article	36	na	na	yes	yes
Silver (1999) Questions (English)	Dissertation	32	yes	na***	yes	yes
Toth (2008) Antiaccusative “se” (Spanish)	Article	78	yes	yes	yes	yes
Ueno (2005) “Te-iru” Construction (Japanese)	Article	44	yes	yes	yes	yes

\* This study did not have a pretest but used a so-called custom-designed posttest that was based exclusively on each learner’s errors made during the task-based interactional treatment. Therefore, zero prior knowledge was assumed for the purposes of calculating the standardized-mean-gain effect size.

\*\* Pretest was administered but the scores were reported for all groups together.

\*\*\* The pretest was administered but consisted of a meta-linguistic judgment and a selected-response components that, by the author’s assertion, turned out to be inadequate measures of target structure development.

Out of the 22 target structures present in the 15 included studies, only 14 target structures appeared in studies whose designs allowed the calculation of the standardized-mean-difference effect size ( $g+$ ) associated with the comparison between the performance of the experimental groups and the control groups on immediate posttests. Therefore, 14 effect-size values were used in the calculation of the weighted mean effect size for the standardized-mean difference on immediate posttests. This number was 10 for delayed posttests. For the standardized-mean-gain effect size, that is, the comparison between the experimental group's performance between the pretest and the immediate and delayed posttest, the numbers of qualifying effect sizes were 18 and 14, respectively.

The weighted standardized-mean-difference effect size ( $g+$ ) for the 14 qualifying effect sizes and the weighted standardized-mean gain for the 18 qualifying effect sizes are positive and show medium and large effects for task-based interaction, respectively. Additionally, the contrasts between the performance of the experimental and comparison groups for the subset of studies that featured a comparison group favored task-based interaction over other types of instructional classroom activities targeting acquisition of grammar. The associated 95% confidence intervals for the standardized-mean difference and the standardized-mean gain did not include zero; therefore, it can be assumed that the overall effect size for task-based interaction was not zero.

The  $Q$  statistic was used to test for homogeneity in the distribution of effect sizes. The chi-square table was used to determine what critical value was needed for statistical significance at the .05 probability level with  $k - 1$  degrees of freedom ( $df$ ), where  $k$  equals the number of studies. The  $Q$  value of 57.07 (for standardized-mean-difference on immediate posttests) exceeded the critical value; therefore, the significant  $Q$  value

indicates heterogeneity of effect sizes. The statistically significant test of homogeneity indicates that the overall mean effect size cannot be assumed to be based on the effect sizes calculated from the 15 included studies. The results of the homogeneity test and the meta-analyst's attempts to find homogeneous sets of effect sizes within the total set are provided in more detail in the *Test of Homogeneity* subsection. In line with Lipsey and Wilson's (2001) recommendation for dealing with heterogeneous sets of effect sizes and the research purpose of the present meta-analysis, a more detailed analysis was conducted to investigate the potential effects of moderator variables that contribute to the heterogeneity of the individual effect sizes.

This chapter first provides a research synthesis of the included studies that, according to Chaudron (2006) and R. Ellis (2006), should be an important integral part of every meta-analysis in the field of second language acquisition (SLA) and cannot be neglected in favor of a purely statistical discussion. Following the section titled *Research Synthesis*, the results of the data analysis are presented in the *Quantitative Meta-Analytic Findings* section by research question, that is, a more general analysis is followed by a differentiated analysis associated with important methodological and pedagogical variables that were represented across the included studies. The analog to the analysis of one-way variance (ANOVA) was used to investigate whether the different levels of specific moderator variables could account for the variability in the effect sizes across the included primary studies.

### Research Synthesis

In this section, a descriptive synthesis of a number of features of the included studies is presented in order to provide an overall picture of the existing research into the

effectiveness of task-based interaction. Study identification features (i.e., the source of the study and the year of publication), some methodological features such as research design, educational setting, and the TL, as well as learner characteristics such as the proficiency level, are tallied and compared across the included study reports. The research synthesis provides a context for interpreting the study results and the basis for formulating recommendations for primary researchers that are presented in chapter V.

### *Research Publication*

The 15 studies that qualified for inclusion in this meta-analysis were based on the inclusion and exclusion criteria specified in chapter III and are marked with an asterisk in the *References* section. Among them, seven studies (46.67%) were published in refereed journals such as *Studies in Second Language Acquisition* ( $k = 4$ ), *Language Awareness* ( $k = 1$ ), *Language Learning* ( $k = 1$ ), and *Japanese Language and Literature* ( $k = 1$ ), whereas one study (6.67%) appeared as a chapter in an edited volume (Adams, 2007; see Table 5). The remaining studies (46.67%) were doctoral dissertations ( $k = 7$ ), three of which were completed at Georgetown University. Some of the dissertation study reports also were published in academic journals (e.g., Revesz, 2006), and, conversely, some of the included journal articles were based on doctoral dissertations (e.g., Iwashita, 2003; Toth, 2008; Ueno, 2005). In such cases, the dissertations were used if available because the dissertations provided more details than the study reports published in journals. No other types of unpublished studies besides doctoral dissertations (e.g., conference reports) that met the criteria and provided sufficient information in order to be included in this meta-analysis were located. Figure 1 shows the publication frequency of included research studies for each year of publication.

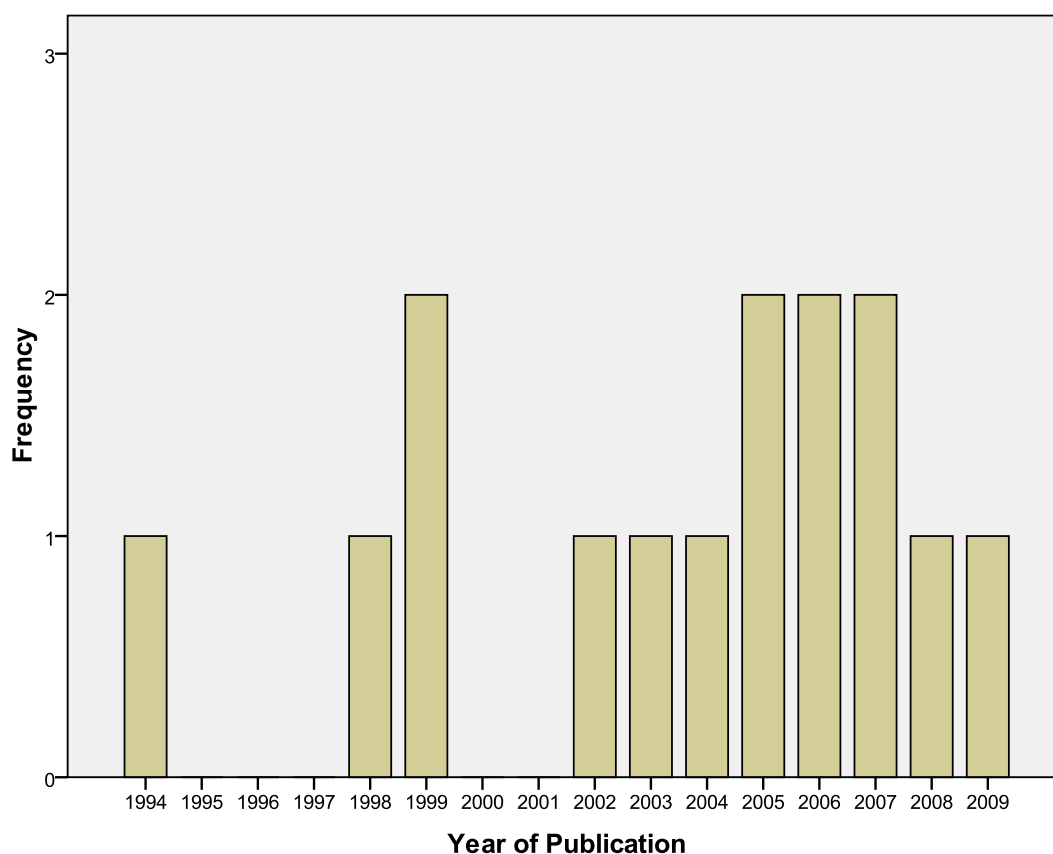


Figure 1. Number of included studies by year of publication.

Even though empirical studies were searched starting with publication year 1980, all included studies fall between the years of 1994 and 2009. Many of the earlier interaction-based studies, especially in the early 1980s, were descriptive rather than experimental or quasi-experimental and were limited to conversational analysis (i.e., analysis of the interlocutors' utterances; Spada & Lightbown, 2009).

#### *Research Setting and Context*

The studies included in this meta-analysis were conducted in a variety of educational settings. The majority of the studies, 66.67% ( $k = 10$ ), were carried out in FL contexts and the remaining 33.33% ( $k = 5$ ) in L2 contexts. Table 6 shows the TL,

language setting (FL or L2), and country where the research study was conducted for all included primary studies.

Table 6

Research Context, Target Language (TL), and Language Setting  
in Included Primary Studies

Study	Target Language	Language Setting	Country
Adams (2007)	English	L2	US
Gass & Alvarez-Torres (2005)	Spanish	FL	US
Horibe (2002)	Japanese	FL	US
Iwashita (2003)	Japanese	FL	Australia
Jeon (2004)	Korean	FL	US
Kim (2009)	English	FL	South Korea
Koyanagi (1998)	Japanese	FL	US
Loschky (1994)	Japanese	FL	US
Mackey (1999)	English	L2	Australia
Nuevo (2006)	English	L2	US
Revesz & Han (2006)	English	L2	US
Revesz (2007)	English	FL	Hungary
Silver (1999)	English	L2	US
Toth (2008)	Spanish	FL	US
Ueno (2005)	Japanese	FL	US

These percentages are similar to the ones reported by Mackey and Goo (2007) and Keck et al. (2006), where 71% of the included studies involved an FL for each of these two meta-analyses. The FL studies included in the meta-analysis involved the following TLs: Japanese ( $k = 5$ ) taught in the US and Australia, Spanish ( $k = 2$ ) taught in the US, English ( $k = 2$ ) taught in Hungary and South Korea, and Korean ( $k = 1$ ) taught in the US. L2 studies involved English ( $k = 5$ ) taught in the US and Australia. Figure 2 shows the TL distribution in the included studies.

Regarding the conditions under which the participants in the treatment groups received instruction, 66.67% of the included studies were laboratory-based ( $k = 10$ ) rather than classroom-based ( $k = 4$ ), and, in one of the studies, some of the participants



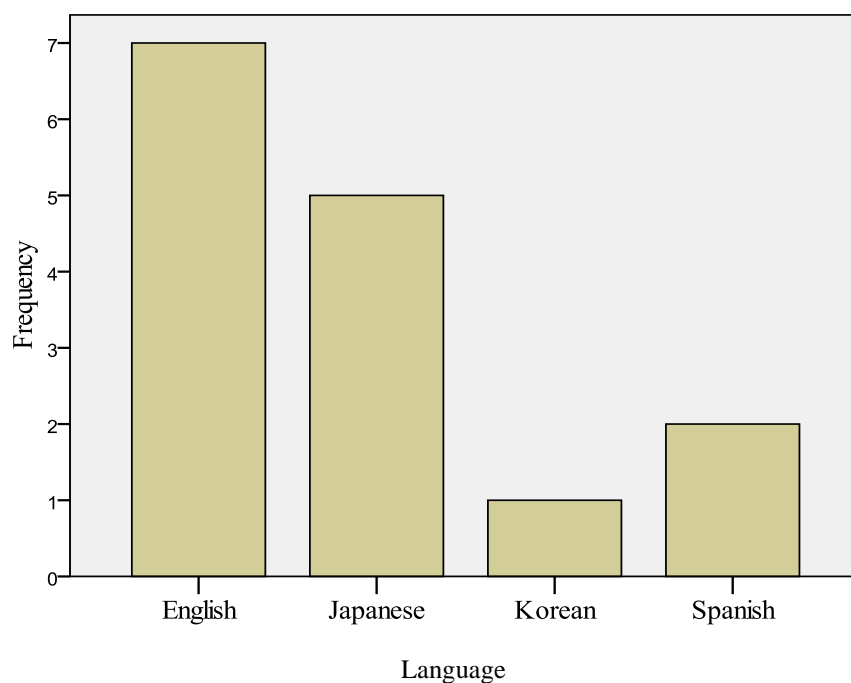


Figure 2. Frequency count for target languages (TLs) in included primary studies.

received one-on-one instruction from the researcher, whereas those participants who were enrolled in the researcher's class received instruction in a regular classroom setting instead. Treatment in laboratory-based studies was provided by native speaker (NS) interlocutors to learners in a one-on-one setting.

### *Learner Characteristics*

Coded learner characteristics included such variables as the participants' first language (L1), age, gender, and TL proficiency level. The majority of the studies involved participants who were university students ( $k = 9$ ). These typically were undergraduate students; however, some of these studies included a mixture of graduate and undergraduate students (e.g., Ueno, 2005). The remaining studies involved adult participants in language courses at US community centers ( $k = 3$ ), a private language

school in Australia ( $k = 1$ ; Mackey, 1999), an intensive language program (IEP) in the US ( $k = 1$ ; Silver, 1999), and high-school students in Hungary ( $k = 1$ ; Revesz, 2007).

The mean age of the participants who were university students based on three of the studies that reported mean age ranged from 18.83 to 20.8 years. Most of the studies, however, reported the age range rather than the mean age, and, for university students, the lower limit was 17 and the upper limit was 36 year old across the studies. The mean age for the participants in adult educational settings such as community centers noticeably was greater, for example, 34.8 years in Revesz and Hans' study (2006; range 20 to 46 years old), 35 years in Adams's study (2007), and in Nuevo's study (2006), the mean age was 33 for the control and the high-complexity-task group and 30 for the low-complexity-task group (range 18 to 62 years). The number of female participants was greater than the number of male participants by 9.00 to 260.00% in eight of the studies that reported these data ( $k = 11$ ), whereas the remaining studies ( $k = 4$ ) did not provide any information about the participants' gender.

The proficiency levels ranged from beginner to high-intermediate and even advanced (for some of the participants in the study) across the included studies; however, the majority of the studies involved either beginners or participants of mixed levels that included beginners as one of the levels ( $k = 13$ ). The institutional course enrollment was the most common way of determining L2 proficiency level, although, in some studies, tests were administered to confirm the participants' proficiency levels, for example, the Australian Second Language Proficiency Rating Scale (Mackey, 1999), Test of English for International Communication (TOEIC) Bridge (Kim, 2009) as well as institutional placement tests (e.g., Toth, 2008) and other departmental tests (e.g., Adams, 2007).

In general, as Keck et al. (2006) and Norris and Ortega (2000) pointed out, the research domain lacks consistent criteria for interpretation of proficiency levels; therefore, different researchers may assign different meanings to such proficiency labels as “beginner” or “intermediate.” Under these circumstances, it was not possible to provide a definitive generalization regarding the participants’ TL proficiency levels across the included studies. Additionally, as mentioned earlier, in some of the studies, (e.g., Koyanagi, 1998), participating learners represented a range of proficiency levels. Finally, in some study reports, learners were classified according to developmental stages in acquisition of specific widely-researched target structures such as, for example, English questions (e.g., Mackey, 1999).

#### *Methodological Features*

There was a great degree of variety in the designs of included studies (see Table 5). Eleven of the 15 included studies (73.33%) used a true control group that did not receive any instruction in the target structure. Two of the remaining four studies used a comparison group, and 6 of the 11 studies with a control group were determined by the meta-analyst to have a comparison group as well. For the purposes of this meta-analysis, all groups that received task-based interaction as the treatment were labeled experimental groups, and any differences between the task-based interaction treatments received by these groups (e.g., task complexity) or additional elements of instruction received (e.g., input that preceded or followed interaction) were treated as potential moderator variables.

The groups that received treatments other than task-based interaction in focused oral communication tasks as defined in this study (e.g., those that received input processing activities or traditional drills) were considered to be comparison groups for the

purposes of this meta-analysis, even though the primary researchers may have referred to them as experimental (i.e., treatment) groups in accordance with their own research purposes. For example, Toth (2008) considered both his learner-led interaction and teacher-led interaction groups experimental, whereas the meta-analyst and the second rater labeled the teacher-led interaction group a comparison group because the manner in which classroom activities were conducted with this large group (approximately 14 participants) did not meet the criteria for focused oral-communication tasks that occur in dyads or small groups specified for the present study.

The number of groups labeled as experimental groups ranged from one to four per study, and the number of comparison groups ranged from one to two. Sample sizes across studies ( $n$ ) ranged from 25 to 191 ( $M = 61.73$ ). The experimental group sample sizes ranged from 7 to 51 participants ( $M = 21.10$ ). The experimental, control, and comparison groups that were present in each study are listed in Table 7. For Jeon's (2004) study, only the numbers of participants that were involved in investigating acquisition of the grammatical target structures out of the total number of participants are provided. (Jeon's study also investigated acquisition of lexis, and the numbers of participants were different for various acquisition targets.)

The majority of the studies used either intact classes ( $k = 8$ ; 53.33%; e.g., Kim, 2009; Toth, 2008) or volunteers ( $k = 5$ ; 33.33%; e.g., Jeon, 2004; Silver, 1999); volunteers were paid in at least one of these studies. Random selection (of 34 participants from the 147 enrolled students in lower proficiency classes) was reported in only one study (Mackey, 1999; 6.67%). It was not possible to determine the basis for participant recruitment in two of the studies (13.33%). One of these studies did not provide any

Table 7  
Study Design and Number of Participants in Included Studies

Study/Group	Number of participants	Total in study
Adams (2007)		
Experimental	25	25
Gass & Alvarez-Torres (2005)		
Experimental, Interaction Only	26	
Experimental, Input + Interaction	19	
Experimental, Interaction + Input	18	
Control	16	
Comparison, Input Only	23	102
Horibe (2002)		
Experimental, Input-Output	11	
Comparison, Input	9	
Control	10	30
Iwashita (2003)		
Experimental	41	
Control	14	55
Jeon (2004)		
Experimental for Honorifics	25 (out of total number)*	
(Experimental for Relative Clauses)	15 (out of total number)*	
Control for Honorifics	9 (out of total number)*	
(Control for Relative Clauses)	6 (out of total number)*	34
Kim (2009)		
Experimental, Simple Task	45	
Experimental, +Complex	47	
Experimental, ++Complex	51	
Comparison, Traditional Instruction	48	191
Koyanagi (1998)		
Experimental, Output	8	
Control	7	
Comparison, Input	8	
Comparison, Output	8	31
Loschky (1994)		
Experimental, Negotiated Interaction	13	
Comparison, Unmodified Input	14	
Comparison, Premodified Input	14	41
Mackey (1999)		
Experimental, Interactor "Readies"	7	
Experimental, Interactor "Unreadies"	7	
Control	7	
Comparison, Scripted Input	6	
Comparison, Observers	7	34
Nuevo (2006)		
Experimental, Low Complexity Task	41	
Experimental, High Complexity Task	32	
Control	30	103

continued on the next page

Table 7 continued

Study/Group	Number of participants	Total in study
Revesz & Han (2006)		
Experimental, Same Video Group	9	
Experimental, Different Video Group	9	
Experimental, Same Notes Group	9	
Experimental, Different Notes Group	9	36
Revesz (2007)		
Experimental, +Photo +Recast	18	
Experimental, –Photo +Recast	18	
Experimental, +Photo –Recast	18	
Experimental, –Photo –Recast	18	
Control	18	90
Silver (1999)		
Experimental, Negotiation	8	
Experimental, “Bare Bones” (Role-Plays)	8	
Control	7	
Comparison, Input Processing	9	32
Toth (2008)		
Experimental, Learner-Led Interaction	25	
Control	25	
Comparison, Teacher-Led Interaction (Non-Task)	28	78
Ueno (2005)		
Experimental	32	
Control	12	44

\* The groups overlapped, and only the participants who received below a certain score for a specific target structure on the pretest were included in the experimental group for that target structure.

information about recruitment at all, and the other study reported that the participants “were chosen” from a certain level; however, it was not clear on what basis they were selected. In terms of participants’ assignment to a specific group (i.e., experimental, control, or comparison), eight of the 15 studies (53.33%; e.g., Iwashita, 2003; Revesz & Han, 2006) utilized random assignment, three studies (20.00%) used statistical control to balance groups for such variables as length of TL study or length of time spent in the target culture, two (13.33%) assigned intact classes to groups, and two (13.33%) did not report the basis for assignment to groups. Just as reported by Keck et al. (2006), none of the studies utilized random sampling. Nevertheless, the percentage of studies using intact

classes and nonrandom assignment was lower than the 70.00% reported by Keck et al. In studies that used intact classes, some efforts to control for confounding variables were reported; for example, each of Toth's (2008) groups (control, comparison, and experimental) consisted of two intact classes taught by different instructors in order to control for quality of instruction and rapport with the participants.

Contrary to the trend reported in previous meta-analyses (Plonsky, 2010), all of the included studies, except for Adams (2007), reported that learners had been given a pretest. Adams used a custom-made posttest that included the items in which learners had made errors during interaction, and therefore their previous competence with these items was assumed to be zero. Additionally, contrary to Keck et al.'s (2006) finding that 57.00% of the studies did not include the description of the pretest that was used, all 14 studies that had a pretest in the present meta-analysis provided such a description. These indicators suggest that the research and reporting practices are improving in the domain.

All included studies in some way investigated the effects of interaction that occurred in focused tasks or the effects of varying oral-communication-task complexity on acquisition of specific target structures as one of their research goals. For example, Koyanagi's (1998) purpose was to investigate the effects of Focus on Form (FoF) tasks on the acquisition of the Japanese conditional "to," whereas Mackey's (1999) main focus was on investigating the effects of ordering of input and interaction (i.e., whether interaction preceding input or interaction following input was more effective). Most of the studies had additional research questions, for example, the role of pair grouping, that is, of being paired with a higher- versus a lower-level proficiency partner (Kim, 2009) or the role of learner differences (field independence vs. field dependence; Ueno, 2005).

Iwashita (2003), among others, investigated the relative impact of using various types of interactional moves produced by NS interlocutors on the development of target structures in the interlanguage of nonnative speakers (NNS). Some of the studies included a qualitative research component; for example, Horibe (2002) investigated how opportunities for spoken output trigger learners' cognitive processes and Iwashita (2003) examined how NS interlocutors respond to nontargetlike utterances produced by NNS interlocutors.

The dependent variable in the included studies typically was interaction-driven morphosyntactic TL development operationalized as improvement in the learners' ability to use the target structures as reflected in their posttest scores. Four of the studies used stage development as the basis for identifying changes in the learners' interlanguage (Adams, 2007; Kim, 2009; Mackey, 1999; Silver, 1999). These studies were based on the developmental framework for English question formation proposed by Pienemann and Johnston (1987) and operationalized the dependent variable as advances in movement through the stage sequence (i.e., stage increase). The following section presents various types of tests that were used to measure participants' acquisition, or development, of target structures.

### *Outcome Measures*

The majority of the included studies employed a pretest, posttest, and a delayed posttest. Out of the 15 included studies, 14 studies (93.33%) utilized a pretest-posttest design (see Table 5). Only Adams (2007) did not use a pretest because she used a custom-made posttest that assumed zero initial ability to use the target structure because it was based on the errors made by individual learners during completion of the treatment



tasks. Loschky (1994) used a pretest; however, he reported results for all three groups used in his study together (i.e., the experimental interaction group as well as the groups that received premodified input and input without interaction that were determined to be comparison groups for the purposes of the present meta-analysis). Therefore, Loschky's results could not be used for combining and comparing standardized-mean-gain effect sizes.

Mackey (1999) had three posttests altogether: an immediate posttest, a second posttest one week later, and a third one 3 weeks later. She did not report separate results for the three posttests but rather the number of learners with a "sustained" stage increase in the target structure development. Therefore, Mackey's results were interpreted as applicable to the final (third) posttest administered 4 weeks after the end of the treatment (delayed posttest). Adams (2007) and Kim (2009) did not have a nonimmediate posttest; the posttests in these studies were administered after 5 and 7 days, respectively.

In the studies that used a pretest, its format was the same as the format of the posttest and the delayed posttest (if the latter was present). All posttests appeared to be researcher-designed except for Silver (1999) who used two forms of the oral-production test that was available commercially from the Language Acquisition Research Center (LARC) at the University of Sydney; however, the researcher also created six additional forms of this test herself. Delayed posttests were administered in 12 studies (80.00%). Three studies (Adams, 2007; Gass & Alvarez-Torres, 2005; Loschky, 1994) did not include a delayed posttest. In Ueno's (2005) study, the control group did not take a delayed posttest so the delayed-posttest scores obtained by the experimental group could only be used to calculate the within-group (i.e., standardized-mean-gain) effect size but

not the between-group (i.e., standardized-mean-difference) effect size.

There was a great deal of variation in the length of time elapsed before the administration of the delayed posttest ( $M = 33.42$ ; range from 7 to 120 days). Keck et al. (2006) and Mackey and Goo (2007) classified posttests with a delay of 0 to 29 days as short-delay, and those with a delay of 30 days or more were labeled long-delay posttests. In the present meta-analysis, this classification would have resulted in only three studies being classified as including a long-delay posttest; therefore, the previous meta-analysts' classification was adjusted slightly. Posttests with a delay of 0 to 27 days ( $k = 6$ ) were considered short-delay posttests, and those with a delay of 28 to 120 days ( $k = 6$ ) were considered to be long-delay posttests.

The classification of outcome measures used in the present meta-analysis was adapted from Norris and Ortega (2000) with an addition of the outcome measure labeled oral-communication as described in chapter III under *Measures of Acquisition of Target Grammatical Structures*. The number of distinct types of outcome measures used within one study (based on this classification) varied between one (e.g., Gass & Alvarez-Torres, 2005; Jeon, 2004) and four (Horibe, 2002). Ueno (2005) reported using more than one type of outcome measure; however, only the total test scores were reported in the journal article. In the remaining 14 studies, out of the five types of outcome measures defined for this meta-analysis, the most frequently utilized type was oral-communication task ( $k = 9$ ; 73.33%), which is a welcome development toward using more communicative forms of assessment that are congruent with the task-based interaction treatments. Metalinguistic judgment was the second most frequently used type of outcome measure ( $k = 7$ ; 46.67%). The frequencies for various types of outcome measures employed in individual studies

based on the classification used in this meta-analysis are presented in Table 8 (oral-communication task is referred to as “Communication Task”).

The tests in Horibe (2002), Koyanagi (1998), and Loschky (1994) had a listening comprehension component that was categorized as a selected-response test based on its format. The free-constructed response in Koyanagi’s (1998) and Revesz’ (2007) studies included both a written and an oral component so the mean effect size was computed for the two components in order to report one effect-size value for this category of outcome measure. Silver (1999) had additional outcome measures (i.e., metalinguistic judgment and selected response) besides the oral communication task; however, this researcher reported that these components of the test did not prove to be good measures of the acquisition of the target structure for various reasons. Therefore, the results of these components of the tests were not used in the analysis.

Norris and Ortega (2000) found that only 16.00% of the studies of effectiveness of L2 instruction that they had reviewed attempted to report any information on the reliability of the outcome measures. Among the research studies included in the present meta-analysis, 73.33% ( $k = 11$ ) reported some information regarding reliability (interrater reliability, internal consistency, and form reliability). This finding constitutes a positive development away from the past trend pointed out by previous meta-analysts (Norris & Ortega, 2000; Russell & Spada, 2006). In fact, in his meta-analysis of interaction-based research completed in 2010, Plonsky (2010) reported that 64.00% of the included study reports contained reliability information.

In regard to instrument validity, many of the primary researchers cited the fact that the outcome measures employed in the studies typically were used as classroom

Table 8  
Types of Outcome Measures Used in Included Studies

Study	Metalinguistic Judgment	Selected Response	Constrained Response	Free Response	Communication Task
Adams (2007)	yes	na	na	na	na
Gass & Alvarez-Torres (2005)	yes	na	na	na	yes
Horibe (2002)	na	yes*	yes	yes	na
Iwashita (2003)	na	na	na	na	yes
Jeon (2004)	na	na	na	na	yes
Kim (2009)	yes	na	na	na	yes
Koyanagi (1998)	yes	yes*	na	yes**	yes
Loschky (1994)	na	yes*	na	na	na
Mackey (1999)	na	na	na	na	yes
Nuevo (2006)	yes	na	na	na	yes
Revesz & Han (2006)	na	na	yes	yes	yes
Revesz (2007)	yes	na	yes	yes**	na
Silver (1999)	na***	na***	na	na	yes
Toth (2008)	yes	na	na	yes	na
Ueno (2005)	Only total score reported				

\* Listening comprehension test

\*\* Included both an oral and written component so the mean effect size was computed

\*\*\* Was present but the test results were discarded based on the primary researcher's assertion that this test was not found to be a good measure of acquisition

tasks or tests (Horibe, 2002) or that similar measures had been used in previous research (e.g., Kim, 2009; Mackey, 1999). In some instances, the tests had been piloted previously on NNS (e.g., Loschky, 1994) or NS (e.g., Revesz, 2007) participants to establish that these tests indeed elicited the use of the target structure. Several primary researchers mentioned other attempts to increase validity and reliability of the outcome measures, for example, by taking steps to ensure that the TL vocabulary that appeared in the instruments did not represent a difficulty for the participants (e.g., Horibe, 2002; Kim, 2009). The impact of the type of test used as the outcome measure on the effect size is discussed in the section titled *Effects of the Type of Outcome Measure*.

#### *Treatment Design and Pedagogical Features*

The duration of the interaction treatment ranged from two sessions (Gass & Alvarez-Torres, 2005) to eight sessions (Ueno, 2005). The total duration of the treatment, therefore, ranged from 45 to 300 minutes. In some instances, the reported time included pretask and posttask activities. Some of the studies, for example, Adams (2007), provided the number of sessions but did not specify the exact duration of the sessions. Other studies provided a range, for example, 15 to 30 minutes for each of the three sessions (Loschky, 1994) and reported deliberately not establishing an upper limit for the interaction in order to make sure that the NNS participants had sufficient time to complete the tasks. For these reasons, any attempts to establish the mean duration of the treatments would be approximate; however, for the purposes of the analysis of the effects of the duration of treatment as a moderator variable, the treatments were divided into “short” (120 minutes or less) and “long” (over 120 minutes; as discussed in the section titled *Effects of the Duration of Treatment*).

The number of different tasks involved in the treatment sometimes reached three of four; however, typically the number of different types of tasks was not greater than two based on the classification presented in chapter II (e.g., information-gap, jigsaw, or role-play tasks). Ueno's (2005) report did not provide sufficient information about the tasks to determine where specifically they would fall in this classification. Silver (1999) had different treatment tasks for the "negotiation" and "bare bones output" groups, both of which were considered experimental in this meta-analysis. (The "bare bones" group completed interactive oral-communication tasks, but the learners did not receive any feedback prompting them to modify their utterances.)

In some instances, it was difficult to determine the task type precisely even when a description was present; for example, in Toth's (2008) study, learners had to sequence a story based on pictures where one partner held all odd-numbered pictures and the other partner held all even-numbered ones. It is hard to say whether a problem-solving component was present in this task, or whether, once the learners completed the information exchange, it was obvious how the pictures were supposed to be sequenced.

Some primary researchers specified the task type in the study report themselves as well as, albeit more rarely, other task characteristics such as whether the task was one-way versus two-way (e.g., Iwashita, 2003), whether the task had one possible outcome (i.e., was a closed task; Loschky, 1994) and whether the participants had the same goal (i.e., convergent task; Loschky, 1994). In the instances where this information was not provided in the primary report, the determination was made by the meta-analyst and the second coder.

Overall, for teaching 7 out of 22 target structures in the included studies

(31.82%), both information-gap and jigsaw tasks together were used. Examples of information-gap tasks were discovering the order of the pictures depicting a story by asking questions (Mackey, 1999) or replicating (i.e., making a drawing of) a picture held by the interlocutor by asking questions about it (Gass & Alvarez-Torres, 2005). An example of a jigsaw task was the most frequently used “spot-the-differences” task (i.e., both interlocutors held pictures and tried to establish what was different between them by asking and answering questions). Additionally, six target structures (27.27%) had treatments that included only jigsaw tasks, and two (9.09%) used information-gap tasks only. Consequently, jigsaw tasks were the most popular type of tasks used in the primary studies included in this meta-analysis as compared with Keck et al.’s (2006) meta-analysis where the most popular type was information-gap tasks: information-gap tasks were used as instructional treatment in eight studies, whereas only one study used a jigsaw task. In the present meta-analysis, in addition to the tasks designed on the information-gap principle, there were two reasoning-gap, specifically, problem-solving, tasks, three information-transfer narrative tasks, and one role-play. Just as in Keck et al.’s meta-analysis, there were no opinion-gap tasks used in the included studies.

The target grammatical structures that were the goal of instruction ranged from one per study (e.g., Japanese conditional “to” in Koyanagi, 1998) to three per study (for example, English questions, past tense, and locative prepositions in Adams, 2007). Overall, there were 22 target structures in the 15 studies. Because some of the target structures were used in more than one study, for example, English past progressive in Revesz (2007) and Revesz and Han (2006) or English questions in Adams (2007), Kim (2009), Mackey (1999), and Silver (1999), only 13 or 14 of these 22 structures were

unique based on the meta-analyst's understanding of their description. (Loschky, 1994 investigated acquisition of two Japanese locative constructions, the results for which were combined, whereas Iwashita, 2003 investigated acquisition of so-called locative-initial constructions.)

In two instances out of 22 (9.09%), the target structures were classified as syntactic, in eight instances as morphological (36.36%), whereas in the remaining 12 instances (54.54%) the structures were deemed to be morphosyntactic (i.e., combining features of morphology and syntax). The Coding Form (see Appendix C) had to be amended to reflect this third category (originally the Coding Form only covered syntactic and morphological structures). For some structures, the classification was provided by the primary researchers, for example, Adams (2007); for others, the determination was made by the meta-analyst and the second coder based on the description of the target structure.

Based on the classification that Spada and Tomita (2010) adopted for their meta-analysis of interactions between the type of instruction and the type of TL feature, in 15 (out of 22) instances (68.18%), structures were determined to be complex (i.e., requiring more than one distinct transformation such as forming most questions in English) and seven (31.82%) were found to be simple (e.g., English past tense such as *washed* or *came*). In the two coders' determination, 17 of 22 target structures (77.27%) in the included studies could be considered relatively unambiguous for learners and five were determined to be ambiguous. These were high-inference decisions that were challenging when the coders were not familiar with the TL. An example of an ambiguous structure is the Japanese "*te -iru*" construction that, according to the primary researcher (Ueno, 2005), expresses the grammatical category of aspect as a temporal property of events and



situations in ways that are unfamiliar to learners whose L1 is English.

Task-essentialness of the target structure was another high-inference coding item. Keck et al. (2006) reported having to make the assumption that the participants used the target structure if its use was intended by task design. A desirable development identified in the present meta-analysis, however, was that many of the primary researchers audio-recorded and subsequently transcribed the interaction. In doing so, they sometimes pursued additional research goals unique to their studies (e.g., Iwashita, 2003; Jeon, 2004; Mackey, 1999); however, in the process, they obtained evidence that the participants indeed used the target structure. Some of the primary researchers provided their own determination regarding the degree of task-essentialness of the target structure (e.g., Revesz & Han, 2006). In some instances, tasks were piloted with NSs, and evidence of task validity in regard to the need for target structure use was obtained in this manner (e.g., Iwashita, 2003; Revesz & Han, 2006). Tasks used by Mackey (1999) had been empirically tested with language learners in previous studies to ensure that they indeed elicited the target structures, and, according to the researcher, previous research had shown that questions could be elicited readily through such tasks. Some authors also asserted that the tasks they used had face validity as familiar classroom materials (e.g., Mackey, 1999). In all studies involving NS interlocutors other than the researchers themselves, training was provided to the participating NS interlocutors.

A number of treatment-related variables presented in chapter II (e.g., task complexity, cognitive characteristics of the learners, presence of explicit instruction in conjunction with task-based interaction, etc.) could not be investigated because of the insufficient number of primary studies that reported these variables at all or with

sufficient clarity. Similarly to the trends reported in the previous meta-analyses, effect-size values were not reported by the primary researchers and, therefore, had to be calculated by the meta-analyst. The considerations that went into the calculations as well as the most important findings are discussed in the following section titled *Quantitative Meta-Analytic Findings*.

### Quantitative Meta-Analytic Findings

This section presents overall weighted mean effect sizes and effect sizes for subgroups of studies sharing various substantive and methodological variables. In addition to the corrected, unbiased mean effect sizes (Hedges's  $g$ ), 95% confidence intervals are presented for each category in which the studies were combined to demonstrate statistical trustworthiness of the reported mean effect sizes (Lipsey & Wilson, 2001). The observed effect is considered to be more robust when it has a narrower confidence interval. When the confidence interval does not include the zero value, the effect is considered statistically significant (i.e., probabilistically different from no effect; Norris & Ortega, 2000).

Overall, 15 unique study samples contributed effect sizes to the meta-analysis that investigated acquisition of 22 distinct target structures; however, the author of one of the studies was also co-author of another study (Revesz, 2007; Revesz & Han, 2006). By some standards, this situation may be viewed as leading to nonindependence of the two study reports (Lipsey & Wilson, 2001). For the purposes of this meta-analysis, due to the fact that the second study was conducted with an entirely different participant sample in a different location (Hungary vs. the US), these two reports were considered independent. Some of the included dissertation studies (Jeon, 2004; Nuevo, 2006) had the same

advisor (Alison Mackey), whose own study (Mackey, 1999) is also included in this meta-analysis. These studies also were considered to be independent reports because they had been conducted with different samples.

As stated in *Research Synthesis*, none of the included studies reported effect sizes for the treatments. The next section briefly reports on the challenges encountered by the meta-analyst in calculating individual effect sizes. Subsequent sections present the results of combining individual effect sizes for the standardized-mean difference (Research Question 1) and standardized-mean gain (Research Question 2) as well as the results of related statistics such as the test of homogeneity. Finally, the section titled *Effects of Moderator Variables* presents the results of aggregation of the effect sizes for various substantive and methodological variables in connection with Research Questions 3, 4, and 5.

### *Calculating Independent Effect Sizes*

Plonsky and Oswald (forthcoming) reported that, in interaction-based SLA research, the aggregation process presents many challenges. It is common for a single primary study to report multiple data on the same relationship between variables from which effect-size values can be calculated. Moreover, there are frequently complex data dependencies in studies with multiple settings, multiple groups, and multiple time points (e.g., immediate and delayed posttests taken at different times). In the present meta-analysis, such challenges were evident as well.

As outlined in chapter III, for studies that employed more than one treatment group (e.g., Kim, 2009; Nuevo, 2006) or more than one group labeled as a comparison group for the purposes of this meta-analysis (Koyanagi, 1998; Loschky, 1994), pooled

means were calculated across all treatment or all comparison groups and used in the effect-size calculation. An additional challenge was that, as presented in *Research Synthesis* in this chapter, studies varied widely in terms of the tests that were used. For example, Revesz (2007) conducted two different oral posttests (with and without the visual support of a photo). The test scores on these two tests were nonindependent (i.e., came from the same group of students); therefore, the effect sizes for each test were calculated separately, and then the mean effect size for the two was computed.

In general, where necessary, effect sizes were aggregated within the study for different types of outcome measures according to the classification presented in chapter II and in *Research Synthesis* in this chapter. For example, both the oral and the written test in Koyanagi's (1998) study were considered to belong to the free-constructed-response type of outcome measure; therefore, one weighted mean effect size was calculated based on the effect sizes associated with these tests. Numerous similar challenges presented themselves in other included primary studies.

In this meta-analysis, one effect size per target structure per study was calculated in compliance with the established meta-analytic practice in the field of SLA where individual structures are believed to have drastic differences from each other (Norris & Ortega, 2006; Spada & Tomita, 2010). The next two sections titled *Standardized-Mean-Difference Effect Size* and *Standardized-Mean-Gain Effect Size* present the results of the aggregation of these single effect-size values obtained for individual target structures for the included studies.

#### *Standardized-Mean-Difference Effect Size*

This section addresses Research Question 1: "To what extent is oral task-based

interaction that occurs in focused (structure-based) communication tasks (in FL and L2 instruction of adult learners) effective (i.e., how large is the standardized-mean-difference effect size resulting from task-based interaction treatments compared with other types of grammar instruction for the learners' acquisition of the target grammatical structure)?” This section mostly focuses on the results of the investigation of the contrasts between the experimental groups that received task-based interaction treatments and the control groups; however, the results of the investigations of the experimental-comparison group contrasts and comparison-control group contrasts also are presented.

The mean between-group, standardized-mean-difference effect size observed across all included studies  $g+ = 0.67$  ( $SE = .08$ ) indicates that treatment groups (i.e., groups that received task-based interaction in focused oral-communication tasks used as the instructional treatment) differed from control groups by approximately two-thirds of a standard-deviation unit on immediate posttests. In congruence with Keck et al.'s (2006) meta-analytic finding, however, the standard deviation for the mean effect size was large ( $SD = .87$ ). (In general, only  $SE$  values are presented in this meta-analysis; however, standard deviations were calculated in some instances to allow for comparisons with Keck et al.'s [2006] and Mackey and Goo's [2007] meta-analytic findings.) On delayed posttests, the gains made by task-based interaction groups as compared with the gains made by control groups were greater than on immediate posttests:  $g+ = 0.71$  ( $SE = .12$ ,  $SD = .78$ ).

In Cohen's (1977) classification, the effect-size values of  $g+ = 0.67$  (for immediate posttests) and  $g+ = 0.71$  (for delayed posttests) correspond to a medium effect size. The 95% confidence intervals (CI) generated around the effect-size estimates were

found to be between 0.50 (lower CI limit) and 0.83 (upper CI limit) for immediate posttests and between 0.47 and 0.95 for delayed posttests, which are relatively narrow bands and thus represents robust results. These confidence intervals did not contain the value of zero, which means that they are statistically significant at alpha level = .05.

Table 9 shows all effect-size values calculated for eligible individual effect sizes ( $k = 14$  for immediate posttests and  $k = 10$  for delayed posttests) that contributed to the weighted mean standardized-mean-difference effect-size values in the present meta-analysis (i.e., studies that employed a control group and administered a posttest or a delayed posttest, or

Table 9  
Standardized-Mean-Difference Effect Sizes Calculated Based on the Contrasts  
Between Experimental and Control Groups

Study/Target Structure	Immediate Posttest		Delayed Posttest	
	<i>g</i>	SE	<i>g</i>	SE
Gass & Alvarez-Torres (2005)				
Gender Agreement (Spanish)	0.10	.28	-	-
“Estar” + location (Spanish)	0.05	.28	-	-
Horibe (2002)				
Temporal Subordinate Conjunctions (Japanese)	2.20	.55	1.67	.53
Iwashita (2003)				
Locative Constructions (Japanese)	0.83	.32	-	-
Verbal “-te” morpheme	0.70	.23	-	-
Jean (2004)				
Honorifics (Korean)	0.69	.40	0.69	.40
Relative Object Clauses (Korean)	1.57	.55	0.91	.50
Koyanagi (1998)				
Conditional “to” (Japanese)	2.86	.74	2.54	.69
Mackey (1999)				
Questions (English)	-	-	1.37	.51
Nuevo (2006)				
Past Tense (English)	0.08	.26	0.16	.31
Locatives (English)	0.15	.22	-0.05	.27
Revesz (2007)				
Past Progressive (English)	1.40	.28	1.52	.51
Silver (1999)				
Questions (English)	0.30	.46	0.44	.46
Toth (2008)				
Antiaccusative “se” (Spanish)	1.51	.32	0.87	.30
Ueno (2005)				
“Te-iru” Construction (Japanese)	1.35	.37	-	-

both; some of these studies had more than one target structure as described in *Research Synthesis*).

All of the standardized-mean-difference effect sizes for the experimental-control contrasts are positive for immediate posttests, and the range is from 0.05 to 2.86. This latter value that comes from Koyanagi's (1998) study represents a very large effect size equal to almost three standard-deviation units. All of the effect sizes for the delayed posttests are positive as well, with the exception of the effect size for the acquisition of locative prepositions ( $g = -0.05$ ; Nuevo, 2006). The maximum effect size value on delayed posttests is 2.57 and also comes from Koyanagi (1998). Figure 3 shows the distribution of the individual between-group, standardized-mean-difference effect sizes

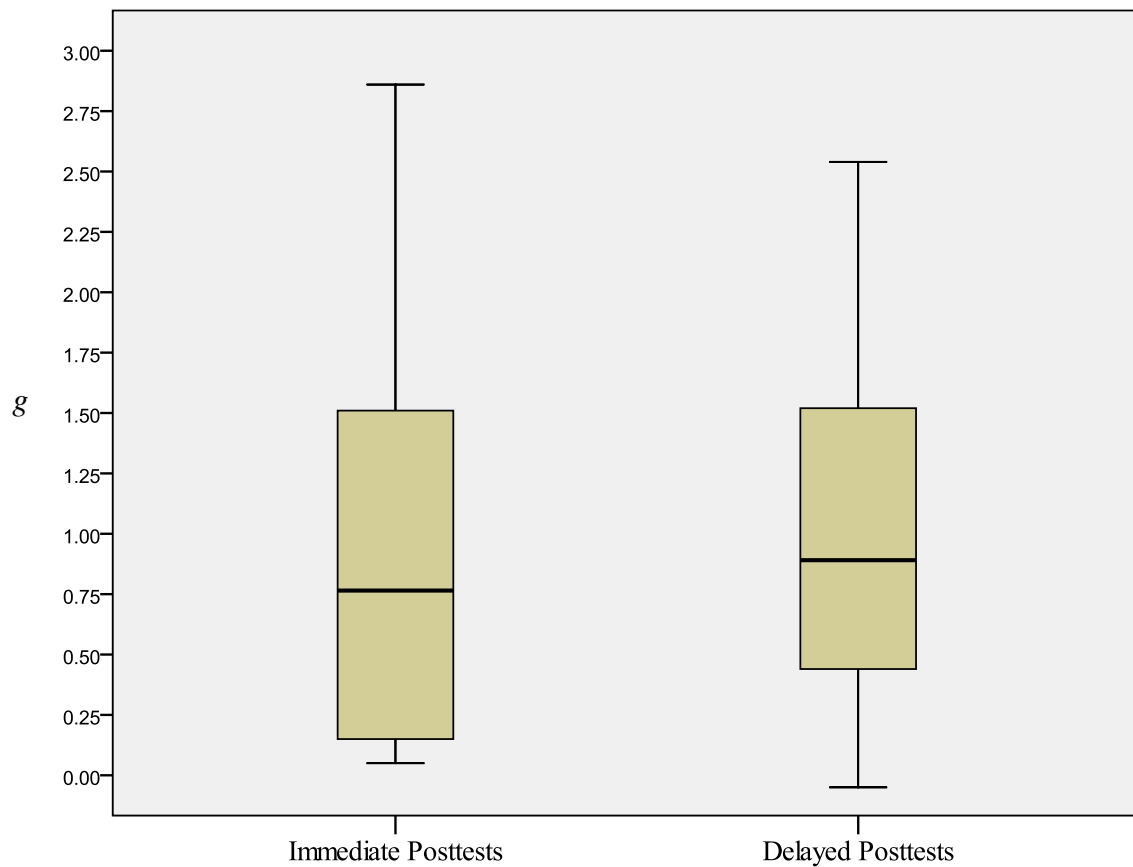


Figure 3. Box plot of standardized-mean-difference effect sizes.

associated with immediate posttests and delayed posttests in a box-and-whisker plot format.

As stated earlier, in SLA research, it is common to consider effect-size values associated with different target structures to be independent from each other even when they come from the same primary study (Keck et al., 2006; Mackey & Goo, 2007; Spada & Tomita, 2010) because different target structures have been shown to be associated with different effect sizes. More stringent requirements, however, demand that only effect-size values from independent reports are included (Cooper, 1998; Lipsey & Wilson, 2001). For this reason, a separate analysis was conducted that included only one randomly selected target structure from the studies that involved more than one target structure (Gass & Alvarez-Torres, 2005; Jeon, 2004; Iwashita, 2003; Nuevo, 2005). The rationale for this decision is presented in more detail in chapter III. The resulting effect-size value  $g+$  was found to be 0.82 for immediate posttests ( $SE = .10$ ,  $SD = .91$ ) and 0.82 ( $SE = .14$ ,  $SD = .80$ ) for delayed posttests, which are large effects according to Cohen (1977). The corresponding CIs were from 0.62 to 1.02 for immediate posttests and 0.55 to 1.10 for delayed posttests. These intervals did not include the value of zero, which means that they also were statistically significant at alpha level = .05. (All subsequent weighted mean effect-size values reported in this meta-analysis were calculated treating effect sizes associated with different target structures as separate, independent values.)

In line with the purpose of the study, the contrasts between the comparison and experimental groups also were investigated. Table 10 shows all effect-size values calculated for eligible individual effect sizes ( $k = 9$  for immediate posttests and  $k = 6$  for delayed posttests) that contributed to the weighted mean standardized-mean-difference



Table 10

Standardized-Mean-Difference Effect Sizes Calculated Based on the Contrasts  
Between Experimental and Comparison Groups

Study/Target Structure	Immediate Posttest		Delayed Posttest	
	g	SE	g	SE
Gass & Alvarez-Torres (2005)				
Gender Agreement (Spanish)	-0.03	.24	-	-
“Estar” + Location (Spanish)	0.11	.28	-	-
Horibe (2005)				
Temp. Subordinate Conjunctions (Japanese)	-0.10	.44	-0.22	.47
Kim (2009)				
Questions	0.55	.17	-	-
Past Tense	0.77	.17	0.68	.17
Koyanagi (1998)				
Conditional “to” (Japanese)	0.75	.45	0.94	.46
Loschky (1994)				
Locative Constructions (Japanese)	-0.21	.34	-	-
Mackey (1999)				
Questions (English)	-	-	0.84	.40
Silver (1999)				
Questions (English)	1.40	.46	1.54	.47
Toth (2008)				
Antiaccusative “se” (Spanish)	-0.55	.28	-0.64	.28

effect-size values in the present meta-analysis based on the contrasts between experimental and comparison groups.

The between-group, standardized-mean-difference effect size for experimental over comparison groups was found to be lower than for experimental-control contrasts ( $g+ = 0.35$ ;  $SE = .09$ ,  $SD = .61$ ; CI from 0.18 to 0.52). According to Cohen’s (1977) classification, the effect size of 0.35 represents a small effect. Nevertheless, this effect-size value still shows that task-based interaction groups on average outperformed groups that received input processing instruction, traditional grammar instruction such as drills and exercises, and so forth. On delayed posttests, the weighted mean effect size for experimental-comparison contrasts even was greater than on immediate posttests:  $g+ = 0.47$  ( $SE = .12$ ,  $SD = .80$ ; CI from 0.22 to 0.70).

Additionally, the weighted mean was calculated separately for the three effect sizes for the experimental groups over those comparison groups that received “traditional” grammar practice and mechanical drills (in Jeon’s [2004] and Koyanagi’s [1998] studies). The resulting effect-size value was  $g+ = 0.68$  ( $SE = .12$ ; CI from .45 to .90). This medium effect size is approximately equivalent to the weighted mean effect size for experimental over control groups.

When compared with control groups, comparison groups showed even larger effects than experimental groups. Table 11 shows all effect-size values calculated for eligible individual effect sizes ( $k = 6$  for immediate posttests and  $k = 5$  for delayed posttests) that contributed to the weighted mean standardized-mean-difference effect-size values in the present meta-analysis based on the contrasts between comparison and control groups.

Table 11

Standardized-Mean-Difference Effect Sizes Calculated Based on the Contrasts  
Between Comparison and Control Groups

Study/Target Structure	Immediate Posttest		Delayed Posttest	
	<i>g</i>	SE	<i>g</i>	SE
Gass & Alvarez-Torres (2005)				
Gender Agreement (Spanish)	0.13	.33	-	-
“Estar” + Location (Spanish)	-0.06	.33	-	-
Horibe (2005)				
Temp. Subordinate Conjunctions (Japanese)	3.23	.68	3.15	.69
Koyanagi (1998)				
Conditional “to” (Japanese)	2.21	.57	1.94	.54
Mackey (1999)				
Questions (English)	-	-	0.54	.48
Silver (1999)				
Questions (English)	-1.07	.54	-1.07	.53
Toth (2008)				
Antiaccusative “se” (Spanish)	1.98	.34	1.60	.32

The weighted mean effect size calculated for the contrasts between the comparison and control groups presented in Table 11 was  $g+ = 0.78$  ( $k = 6$ ,  $SE = .17$ , CI from 0.46 to 1.10) on immediate posttests, which is on the high end of the medium range according to Cohen's (1977) interpretation guidelines, and  $g+ = 1.19$  ( $k = 5$ ,  $SE = .21$ , CI from 0.78 to 1.59) on delayed posttests, which is a large effect (Cohen, 1977). Discussion of the findings presented in this section and their implications are presented in chapter V.

#### *Standardized-Mean-Gain Effect Size*

This section deals with Research Question 2: "Is the standardized-mean-gain effect size (i.e., effect size based on the pre- to posttest differences) larger for task-based interaction treatments as compared with other types of grammar instruction?" This section focuses in greater detail on the pre- to posttest gains made by the experimental groups; however, the results of the investigation of the gains made by the control and comparison groups also are reported for comparison purposes.

The standardized-mean-gain effect size is based on within-group changes (i.e., from the pretest to the immediate or delayed posttest). Keck et al. (2006) and Norris and Ortega (2000) reported calculating the standardized-mean-gain effect size based on the pretest-posttest differences for the small subset of the studies that did not have control or comparison groups, whereas Mackey and Goo (2007) computed this type of effect size for all studies that contained information about pretest to posttest changes (even if a control or comparison group was present).

Plonsky and Oswald (forthcoming) reported that some meta-analyses in the SLA field mistakenly have treated both types of effects as comparable. In general, pretest-posttest contrasts tend to produce larger effects (Morris, 2008), are a different type of

estimation than between-group contrasts, and therefore should be treated separately. It is common in SLA meta-analyses to apply the between-group formula for the  $d$ -value to the within-group, pretest-posttest designs. The problem is that the more appropriate formula (Lipsey & Wilson, 2001, p. 44) requires the correlation ( $r$ ) between pre- and posttest scores. This correlation almost never is reported in primary studies (Plonsky, 2010), and, without it, the resulting effect size will be biased (Cheung & Chan, 2004; Gleser & Olkin, 2009; Plonsky & Oswald, forthcoming). In the present meta-analysis, the two types of effect size estimates (i.e., standardized-mean-difference and standardized-mean-gain) are treated separately. Nevertheless, in the absence of  $r$ -values needed to apply the appropriate formula and in line with the previous practices, the between-group formula was used for within-group contrasts with an understanding that the resulting effect sizes are estimates and may be upwardly biased.

Table 12 shows all effect-size values calculated for individual studies that contributed to the weighted mean standardized-mean-gain value in the present meta-analysis. The calculated weighted mean effect size associated with within-group (i.e., pre- to posttest) comparisons was  $g+ = 1.09$  ( $SE = .06$ ,  $SD = 1.05$ ; CI from 0.97 to 1.20), which means that the experimental groups showed considerable change from the pretest to the immediate posttest after receiving task-based interaction as instructional treatment. The confidence interval was narrow, and this finding was statistically significant at alpha level = .05.

All of the standardized-mean-gain effect sizes are positive for both immediate and delayed posttests. The range for immediate posttests is 0.13 to 3.31. The latter value comes from Adams' (2007) study in which the primary researcher used custom-made posttests where the participants were scored exclusively on the items in which they had

Table 12

Standardized-Mean-Gain Effect Sizes Calculated for Experimental Groups

Study/Target Structure	<u>Immediate Posttest</u>		<u>Delayed Posttest</u>	
	g	SE	g	SE
Adams (2007)				
Questions (English)	2.33	.37	-	-
Past Tense (English)	3.31	.44	-	-
Locatives (English)	2.78	.40	-	-
Gass & Alvarez-Torres (2005)				
Gender Agreement (Spanish)	0.34	.18	-	-
“Estar” + location (Spanish)	0.13	.18	-	-
Horibe (2002)				
Temp. Subordinate Conjunctions (Japanese)	2.35	.55	2.02	.55
Iwashita (2003)				
Locative Constructions (Japanese)	0.73	.23	0.79	.23
Verbal “-te” morpheme	1.13	.24	1.09	.24
Jeon (2004)				
Honorifics (Korean)	0.78	.29	0.76	.29
Relative Object Clauses (Korean)	2.34	.49	1.74	.43
Kim (2009)				
Past Tense (English)	1.57	.14	1.57	.14
Koyanagi (1998)				
Conditional “to” (Japanese)	2.62	.68	2.28	.64
Mackey (1999)				
Questions (English)	-	-	2.13	.47
Nuevo (2006)				
Past Tense (English)	0.39	.19	0.16	.21
Locatives (English)	0.28	.17	0.21	.18
Revesz (2007)				
Past Progressive (English)	1.63	.29	1.70	.23
Revesz & Han (2006)				
Past Progressive (English)	3.11	.35	3.26	.36
Toth (2008)				
Antiaccusative “se” (Spanish)	1.42	.32	0.83	.29
Ueno (2005)				
“Te-iru” Construction (Japanese)	2.56	.34	2.90	.36

made errors during interaction. Therefore, such a large effect-size value can be expected.

In his meta-analysis, Plonsky (2010) considered values greater than 3.00 to be outliers

and excluded them from the analysis; however, in the case of pre- to posttest

comparisons, effect-size values are expected to be greater. Figure 4 shows the distribution

of the within-group, standardized-mean-gain effect sizes associated with immediate

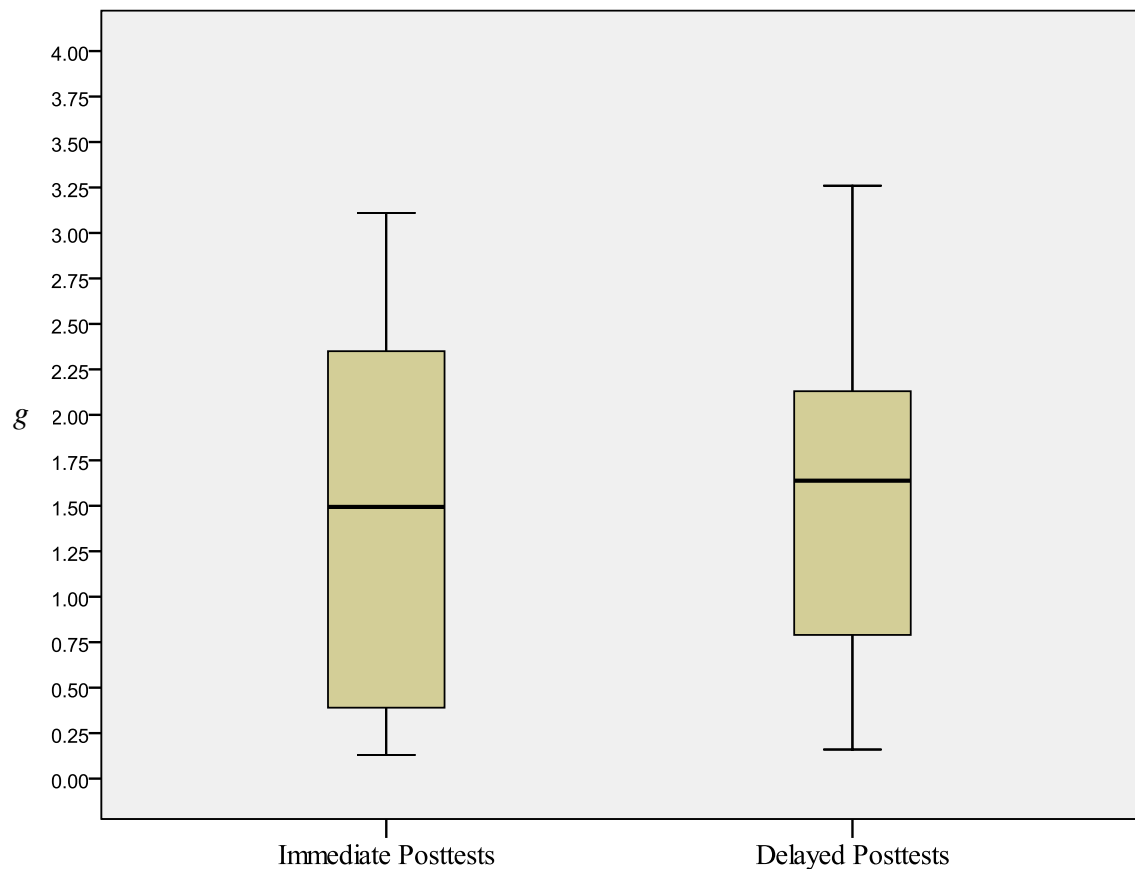


Figure 4. Box plot of standardized-mean-gain effect sizes.

posttests and delayed posttests in a box-plot format.

For acquisition of the four target structures investigated in the two studies that did not have a control or comparison group (and, therefore, could only yield standardized-mean-gain effect size values; Adams, 2007; Revesz & Han, 2006), the weighted mean effect size also was calculated separately:  $g_{+} = 2.89$  ( $SE = .19$ ;  $SD = .43$ , CI from 2.48 to 3.24).

In general, the standardized-mean-gain effect size even was greater on the delayed posttests:  $g_{+} = 1.19$  ( $SE = .07$ ;  $SD = .94$ ; CI from 1.05 to 1.32). The findings for both immediate and delayed posttests represent an increase of more than one standard-

deviation unit and can be interpreted as large effects based on Cohen's (1977) proposed classification. For comparison, control groups showed substantially smaller gains on immediate posttests  $g+ = 0.16$  ( $SE = .10$ ,  $SD = .36$ ; CI from -0.05 to 0.36) and delayed posttests  $g+ = 0.32$  ( $SE = .14$ ,  $SD = .64$ ; CI from -0.05 to 0.36) than treatment groups. (Table 13 shows standardized-mean-gain effect sizes calculated for control groups as well as for comparison groups in individual studies.)

Unlike control groups, comparison groups showed large pre- to posttest gains (like experimental groups). The weighted standardized-mean-gain effect size calculated across the comparison groups was  $g+ = 0.92$  on immediate posttests ( $SE = .13$ ,  $SD = 1.09$ ; CI from 0.67 to 1.17) and  $g+ = 1.22$  on delayed posttests ( $SE = .15$ ,  $SD = .85$ ; CI from 0.93 to 1.50), both of which are large effects based on Cohen's (1977) classification. The discussion of these findings and their implications is presented in chapter V. The next section provides a detailed presentation of the results of the heterogeneity test conducted for the effect sizes discussed in the *Standardized-Mean-Difference Effect Size* and the *Standardized-Mean-Gain Effect Size* sections.

#### *Test of Homogeneity*

The rather high standard deviations for the weighted mean effect sizes reported in the previous two sections indicated that there was a wide degree of dispersion among the effect sizes calculated for individual included studies. The findings of medium and large effects and high standard deviations are consistent with the findings presented in Keck et al.'s (2006) and Mackey and Goo's (2007) meta-analyses.

Hedges's  $Q$  statistic (Hedges & Olkin, 1985) was used to analyze the distribution of standardized-mean-difference and standardized-mean-gain effect sizes presented in

Table 13

Standardized-Mean-Gain Effect Sizes Calculated for Control and Comparison Groups

Study/Target Structure	<u>Immediate Posttest</u>				<u>Delayed Posttest</u>			
	<u>Control</u>		<u>Comparison</u>		<u>Control</u>		<u>Comparison</u>	
	<i>g</i>	SE	<i>g</i>	SE	<i>g</i>	SE	<i>g</i>	SE
Gass & Alvarez-Torres (2005)								
Gender Agreement (Spanish)	-0.03	.35	0.17	.30	-	-	-	-
“Estar” + location (Spanish)	0.26	.36	0.31	.30	-	-	-	-
Horibe (2002)								
Temp. Subord. Conjunctions (Japanese)	0.36	.45	2.74	.62	2.02	.55	2.96	.67
Iwashita (2003)								
Locative Constructions (Japanese)	0.05	.38	-	-	0.80	.46	-	-
Verbal “-te” morpheme (Japanese)	0.82	.39	-	-	1.09	.24	-	-
Jeon (2004)								
Honorifics (Korean)	0.04	.79	-	-	0.03	.79	-	-
Relative Object Clauses (Korean)	0.18	.88	-	-	0.31	.91	-	-
Kim (2009)								
Past Tense (English)	-	-	0.70	.21	-	-	0.71	.21
Koyanagi (1998)								
Conditional “to” (Japanese)	0.77	.55	2.27	.47	2.00	.65	2.06	.45
Mackey (1999)								
Questions (English)	-	-	-	-	0.72	.55	1.29	.43
Nuevo (2006)								
Past Tense (English)	0.48	.31	-	-	0.27	.35	-	-
Locatives (English)	0.03	.26	-	-	0.16	.29	-	-
Revesz (2007)								
Past Progressive (English)	0.23	.33	-	-	-0.09	.41	-	-
Toth (2008)								
Antiaccusative “se” (Spanish)	-0.18	.28	1.90	.32	0.16	.28	1.48	.30
Ueno (2005)								
“Te-iru” Construction (Japanese)	0.47	.41	-	-	2.90	.36	-	-

Table 9 (standardized-mean-difference effect sizes for experimental and control group comparisons) and Table 12 (standardized-mean-gain effect sizes for the experimental groups). The homogeneity test (i.e.,  $Q$  statistic) was used to assess whether the variance in values yielded from the included studies was statistically significantly different from sampling error (Cooper, 1998). The  $Q$  statistic is distributed as a chi-square; therefore,



the chi-square table was used to determine what critical values were needed for statistical significance at the .05 probability level with  $k - 1$  degrees of freedom, where  $k$  equals the number of individual effect sizes included in the calculation of the weighted mean.

The  $Q$  value is computed as the summation over the products of the inverse variance weight for each individual effect size and the squared difference between each individual effect size and the mean effect size. In this analysis,  $Q$  values were calculated for the following: (a) standardized-mean-difference effect sizes on immediate posttests, (b) standardized-mean-difference effect sizes on delayed posttests, (c) standardized-mean-gain effect sizes on immediate posttests, and (d) standardized-mean-gain effect sizes on delayed posttests. All four calculated  $Q$  values were greater than the critical values in the chi-square table for the respective degrees of freedom, which means that all four sets of effect sizes were heterogeneous, rather than homogeneous. Table 14 contains the calculated  $Q$  values for these four sets of effect sizes.

As explained in chapter III, the meta-analyst attempted to find homogeneous sets of effect sizes within these heterogeneous sets by removing outliers, Winsorizing, removing studies with low numbers of participants, and so forth. These attempts largely

Table 14  
Results of the Homogeneity Test ( $Q$  Statistic)

	<u>Effect Size</u>		<u>95% Confidence Interval</u>		<u>Homogeneity of Effect Sizes</u>	
	<u>g+</u>	<u>SE</u>	<u>Lower</u>	<u>Upper</u>	<u>Q value</u>	<u>df</u>
Standardized-mean difference (between groups) on immediate	0.67	.08	0.50	0.83	57.07*	13
Standardized-mean difference (between groups) on delayed	0.71	.12	0.47	0.95	27.63*	9
Standardized-mean gain (within groups) on immediate	1.09	.06	0.97	1.20	228.43*	17
Standardized-mean difference (within groups) on delayed	1.19	.07	1.05	1.32	140.84*	13

\* Statistically significant at .05 level

were unsuccessful. For example, a homogeneous set (at .05 level) could be established for the standardized-mean-difference effect size associated with immediate posttests by either removing the six greatest effect-size values or by removing the five lowest and two greatest values simultaneously (out of 14 values). The task of finding a homogeneous set of effect sizes for the standardized mean gain proved to be even more challenging. With all but the six (out of 18) middle effect sizes removed ased on identified natural “breaking points,” the  $Q$  statistic returned a value that still was larger than the critical value at the .05 level.

Additionally, the face examination of the effect sizes that had to be removed to achieve homogeneity suggested that this procedure possibly duplicated the analysis of moderator variables. For example, the majority of the larger effect sizes appeared to be associated with the studies involving a TL characterized by the greatest distance from the learners’ L1 (i.e., Japanese and Korean when the learners’ L1 is English) as an FL in a university setting (see the *Effects of Moderator Variables* section for the results of the investigation involving these moderator variables). For these reasons and in order to avoid further reduction of the already scarce data, the original heterogeneous sets were retained for the analysis. The decision to use the original heterogeneous sets appeared to be in line with the practice established in the previous meta-analyses in the research domain where the tests of homogeneity typically were not conducted.

According to Lipsey and Wilson (2001), if the homogeneity-of-variance assumption is violated, the meta-analyst may assume that the source of variation potentially is due to some moderating variables. Then the meta-analyst can proceed to examine the effects of these moderator variables such as specific characteristics of the

task-based interaction treatment, study characteristics, learner characteristics, and so forth that have been recorded during the coding process. The next section presents the results of the investigation of these variables.

### *Effects of Moderator Variables*

This section deals with Research Questions 3, 4, and 5. It examines the effects of potential moderator variables such as characteristics of the tasks used as treatment, types of outcome measures used in the study as well as other pedagogical and methodological variables.

#### *Effects of Task Type*

The next two subsections deal with Research Question 3: “Is there a difference in effect size values based on the type of focused communication task (e.g., information-gap vs. opinion-gap, closed vs. open, etc.) used in the task-based interaction treatment”? These subsections present major findings associated with the most important variables representing various task characteristics and the results of the analog to ANOVA used to check whether the levels of these variables can account successfully for the differences in aggregated effect sizes.

*Task type based on the gap principle.* In accordance with the classification of task types based on the so-called gap principle that was provided in the section titled *The Gap Principle and Major Task Designs* in chapter II, the following task types were coded among those associated with the effect sizes used in this meta-analysis: (a) information-gap ( $k = 2$ ), (b) jigsaw ( $k = 6$ ), (c) information-gap and jigsaw ( $k = 7$ ; when both of these task types were used as treatment for the same target structure in one study), (d) problem-solving ( $k = 2$ ), (e) narrative (information-transfer;  $k = 3$ ), and (f), and role-play ( $k = 1$ ).

The studies that involved multiple target structures used either the same or different tasks as treatment for individual target structures. For reasons presented in the *Research Synthesis under Treatment Design and Pedagogical Features*, Silver's (1999) "bare bones output" group and Ueno's (2003) experimental groups were not included in this part of the meta-analysis.

In many instances, there were only one or two effect sizes of a specific type (e.g., standardized-mean-difference on delayed posttests) associated with a particular type of task; therefore, analysis was conducted for only two levels of the task type variable: (a) all information-gap and jigsaw tasks (because both these types fall under the information gap based on the gap-principle classification) and (b) all other types of tasks (i.e., problem-solving, narrative, and role plays). The between-group data for the contrasts between experimental and comparison groups were not included in this part of the analysis because of the insufficient number of qualifying studies for the level labeled "other types of tasks" (i.e.,  $k < 3$ ). Table 15 presents the weighted mean effect-size values (standardized-mean-difference on immediate and delayed posttests and standardized-mean-gain on immediate and delayed posttests), standard error, confidence intervals, and results of the analog to ANOVA for task types classified based on the gap principle. Additionally, Table 15 shows the results of the analysis for one-way versus two-way tasks. Because some of the included studies involved treatments that contained both one-way and two-way tasks (e.g., an information-gap and a jigsaw task in one treatment) some of the treatments were labeled as "mixed" and the numbers of the qualifying effect sizes for some types of comparisons were low. For this reason, only the weighted mean effect sizes for the comparisons between the experimental and control groups on

Table 15

Weighted Mean Effect Sizes for the Variable of Task Type (Based on the Gap Principle)  
and One-way versus Two-way Tasks

Type of Task	Type of Effect Size	k	Effect Size		95% CI		Q <sub>W</sub>	Q <sub>B</sub>
			g <sup>+</sup>	SE	Lower	Upper		
	Between groups:							
IG and Jigsaw	Exp – Control (Immediate)	7	0.71	.12	0.48	0.92	25.62*	0.10
Other	Exp – Control (Immediate)	5	0.69	.13	0.44	0.94	33.29*	
IG and Jigsaw	Exp – Control (Delayed)	6	0.97	.18	0.63	1.31	8.77	3.84
Other	Exp – Control (Delayed)	4	0.48	.18	0.14	0.82	20.24*	
	Within groups:							
IG and Jigsaw	Exp (Pre to Immediate)	11	1.07	.07	0.93	1.21	112.76*	0.10
Other	Exp (Pre to Immediate)	6	0.97	.11	0.77	1.18	95.27*	
IG and Jigsaw	Exp (Pre to Delayed)	7	1.25	.09	1.08	1.43	18.16*	2.29
Other	Exp (Pre to Delayed)	5	0.87	.11	0.65	1.09	86.95*	
	Between groups:							
One-way	Exp – Control (Immediate)	4	0.87	.15	0.59	1.16	5.60	6.03
Two-way	Exp – Control (Immediate)	6	0.59	.13	0.33	0.86	34.69*	
Mixed	Exp – Control (Immediate)	4	0.33	.15	0.00	0.66	7.63	

\* Statistically significant when the overall comparison error rate was controlled at .05 level

immediate posttests are included in Table 15.

As can be seen from the results presented in Table 15, the effect-size values associated with treatments involving jigsaw or information-gap tasks, or both, were greater than the effect-size values associated with the “other” types of tasks; however, the confidence intervals overlapped and the results of the analog to ANOVA statistic were not statistically significant. The overall weighted mean effect size for one-way tasks was greater than the mean effect size for two-way tasks and substantially greater than for treatments that involved both one-way and two-way tasks together (i.e., mixed). Nevertheless, the 95% confidence intervals overlapped for these three levels of this variable and the results of the analog to ANOVA statistic were not statistically significant

for  $Q_B$ . (If the corresponding  $Q_W$  values are not statistically significant and  $Q_B$  is statistically significant, then the variation in the effect sizes can be explained by the levels of the moderator variable.) The next section presents the results associated with other important task characteristics explored as potential moderator variables.

*Open-endedness and convergence.* This section presents the weighted mean effect sizes, standard error, 95% confidence intervals, and the results of the analog to ANOVA statistic for two more important variables associated with task design: (a) open-endedness (closed vs. open tasks based on whether there is only one or more than one possible solutions) and (b) convergence (convergent vs. divergent based on whether the interlocutors have the same or different goals as determined by the task). These task characteristics are described in more detail in chapter II. The results of the analysis for these two variables for the types of comparisons that had at least three associated effect sizes for each level of the variable are presented in Table 16.

As can be seen from Table 16, closed tasks that require the participants to reach one predetermined solution were associated with a medium standardized-mean-difference effect size on immediate posttests  $g+ = 0.70$ , whereas open tasks were associated with a small effect  $g+ = 0.37$ , even though the confidence intervals overlapped.

The standardized-mean-difference effect size associated with divergent tasks on immediate posttests was large  $g+ = 1.45$  and was considerably greater than the small effect size associated with convergent tasks  $g+ = 0.47$ . The confidence intervals did not overlap, which indicates a statistically significant difference. The analog to ANOVA for convergent versus divergent tasks returned a  $Q_B$  value that exceeded the critical value

Table 16

Weighted Mean Effect Sizes for the Variables of Open-endedness and Convergence

Variable	Type of Effect Size	k	Effect Size		95% CI		$Q_W$	$Q_B$
			g+	SE	Lower	Upper		
Open-endedness	Between groups:							
Closed	Exp – Control (Immediate)	9	0.70	.11	0.49	0.91	26.86*	1.93
Open	Exp – Control (Immediate)	3	0.37	.21	-0.06	0.79	12.76*	
	Within groups:							
Closed	Exp (Pre to Immediate)	9	0.98	.07	0.83	1.12	75.33*	5.37
Open	Exp (Pre to Immediate)	4	1.37	.15	1.07	1.67	66.09*	
Closed	Exp (Pre to Delayed)	7	1.37	.09	1.19	1.54	19.71*	1.58
Open	Exp (Pre to Delayed)	4	1.13	.17	0.80	1.45	62.42*	
Convergence	Between groups:							
Convergent	Exp – Control (Immediate)	7	0.47	.11	0.27	0.68	19.12*	28.10*
Divergent	Exp – Control (Immediate)	5	1.54	.19	1.07	1.83	9.26	
Convergent	Exp – Control (Delayed)	5	0.42	.15	0.12	0.71	9.52	11.20
Divergent	Exp – Control (Delayed)	5	1.29	.21	0.87	1.71	6.91	
	Within groups:							
Convergent	Exp (Pre to Immediate)	11	0.90	.06	0.78	1.03	146.16*	32.63*
Divergent	Exp (Pre to Immediate)	6	1.86	.16	1.56	2.17	29.88*	
Convergent	Exp (Pre to Delayed)	8	0.95	.08	0.79	1.10	63.53*	23.77*
Divergent	Exp (Pre to Delayed)	5	1.76	.15	1.47	2.05	29.76*	

\* Statistically significant when the overall comparison error rate was controlled at .05 level

when comparison error rate was controlled at .05 level; however, the  $Q_W$  for convergent tasks was also statistically significant, which did not allow to make the determination that this moderator variable (i.e., convergence) successfully accounted for the variability in effect sizes (Lipsey & Wilson, 2001).

The observed effects also were larger for divergent tasks (large effect) than for convergent tasks (small effect) on delayed posttests on between-group comparisons. Within-group comparisons yielded large effects for both divergent and convergent tasks; however, the effects for divergent tasks were considerably greater and approximated 2 standard deviation units (as shown in Table 16). The results of the analog to ANOVA, however, failed to confirm that this task characteristic (i.e., convergence vs. divergence) accounted for the variability in the effect sizes. Additional moderator variables related to

the characteristics of the target structure that is the focus of instruction are investigated in the next section.

### *Effects of Characteristics of Target Structures*

This section addresses part of Research Question 4: “Is there a difference in effect-size values based on other factors such as the type of grammatical structure targeted by the task-based-interaction treatment, duration of instruction as well as miscellaneous other teacher-related, learner-related, and contextual variables?” In particular, the results associated with the effects of the type of the target structure on the effectiveness of task-based interaction conducted for the purpose of facilitating the learners’ acquisition of this structure are presented.

The following characteristics of the target structures were analyzed: (a) morphological versus morphosyntactic (syntactic structures were not analyzed as a separate level of the variable because of the insufficient number of the studies in which treatment involved syntactic structures), (b) simple versus complex (using Spada & Tomita’s [2010] criteria), and (c) ambiguous versus unambiguous (based on the information provided in the primary studies and the inferences made by the two coders). Table 17 contains the meta-analytic findings for these three variables (when the number of associated effect sizes for each level of the variables was three or greater).

Weighted mean effect sizes for acquisition of morphosyntactic structures were greater than for morphological structures for both between-group and within-group contrasts on immediate and delayed posttests. Additionally, the confidence intervals did not overlap for these effect sizes, except for the experimental-control group contrast on delayed posttests where the confidence interval for morphological structures also



Table 17

## Weighted Mean Effect Sizes Associated with Characteristics of the Target Structures

Variable	Type of Effect Size	k	Effect Size		95% CI		Q <sub>W</sub>	Q <sub>B</sub>
			g+	SE	Lower	Upper		
Morphology/Syntax	Between groups:							
Morphological	Exp – Control (Immed.)	10	0.50	.10	0.31	0.68	34.38*	18.94
Morphosyntactic	Exp – Control (Immed.)	3	1.58	.23	1.13	2.03	3.47	
Morphological	Exp – Control (Delayed)	3	0.34	.18	-0.02	0.70	10.86*	5.67
Morphosyntactic	Exp – Control (Delayed)	6	0.94	.18	0.60	1.29	7.66	
	Within groups:							
Morphological	Exp (Pre to Immediate)	8	0.83	.07	0.69	0.96	110.47*	71.18*
Morphosyntactic	Exp (Pre to Immediate)	8	2.07	.13	1.81	2.33	39.12*	
Morphological	Exp (Pre to Delayed)	5	1.01	.08	0.88	1.17	62.77*	24.24
Morphosyntactic	Exp (Pre to Delayed)	8	1.80	.14	1.53	2.06	50.58*	
Simple/Complex	Between groups:							
Simple	Exp – Control (Post)	4	0.10	.13	-0.15	0.35	0.09	35.42*
Complex	Exp – Control (Post)	10	1.11	.11	0.89	1.34	21.56	
	Within groups:							
Simple	Exp (Pre to Immediate)	7	0.82	.07	0.67	0.96	125.37*	39.52*
Complex	Exp (Pre to Immediate)	11	1.58	.10	1.39	1.77	63.54*	
Simple	Exp (Pre to Delayed)	3	0.88	.10	0.69	1.07	52.02*	20.23
Complex	Exp (Pre to Delayed)	11	1.49	.10	1.30	1.68	68.59*	
Ambiguity	Between groups:							
Unambiguous	Exp – Control (Immed.)	9	0.49	.11	0.29	0.70	19.17	16.59
Ambiguous	Exp – Control (Immed.)	4	0.96	.18	0.62	1.31	21.31*	
Unambiguous	Exp – Control (Delayed)	7	0.59	.15	0.30	0.88	19.70*	2.22
Ambiguous	Exp – Control (Delayed)	3	0.99	.23	0.55	1.43	5.71	
	Within groups:							
Unambiguous	Exp (Pre to Immediate)	13	1.14	.07	1.02	1.27	172.48*	3.85
Ambiguous	Exp (Pre to Immediate)	5	0.87	.13	0.62	1.11	52.10*	
Unambiguous	Exp (Pre to Delayed)	10	1.15	.07	1.00	1.29	111.18*	1.72
Ambiguous	Exp (Pre to Delayed)	4	1.39	.17	1.06	1.73	27.94*	

\* Statistically significant when overall comparison error rate was controlled at .05 level

included zero. The difference in favor of morphosyntactic structures especially was pronounced for standardized-mean gain on immediate posttests:  $g+ = 2.07$  as compared with  $g+ = 0.83$  for morphological structures.

Effect sizes associated with complex target structures were greater than those associated with simple structures. For example, the standardized-mean-difference (i.e., between-group) effect size value for simple structures ( $g+ = 0.10$ ) was smaller substantially than for complex structures ( $g+ = 1.11$ ) on immediate posttests, with nonoverlapping confidence intervals; however, the confidence interval for simple

structures included zero. As shown in Table 17, similar results were obtained for within-group comparisons on both immediate and delayed posttests: the weighted mean effect sizes associated with complex structures exceeded significantly the values associated with simple structures, even though all within-group effect sizes were large based on Cohen's (1977) suggested interpretation guidelines. The confidence intervals did not overlap.

Acquisition of ambiguous structures tended to be associated with greater weighted mean effect sizes than of structures determined to be unambiguous, except in the case of the standardized-mean-gain effect size on immediate posttests (even though the confidence intervals overlapped in all cases). In the latter case, the effect was smaller for ambiguous structures ( $g+ = 0.87$ ) than for unambiguous ( $g+ = 1.14$ ); however, both effects could be interpreted as large based on Cohen's (1977) guidelines.

The analog to ANOVA was performed for all three variables related to the characteristics of the target structures that are presented in this section. The requirements for statistically significant  $Q_B$  values with the associated  $Q_W$  values being nonsignificant were met only for the between-group comparison between simple and complex structures. The next section examines the duration of the task-based-interaction treatment as a potential moderator variable.

#### *Effects of the Duration of Treatment*

This section addresses part of Research Question 4 that has to do with the effects of the duration of task-based interaction treatment received by participants on their acquisition of target structures. The weighted mean effect sizes, standard error, and 95%

confidence intervals for short (i.e., 120 minutes or less) and long (i.e., over 120 minutes) treatments are presented in Table 18.

The weighted standardized-mean-difference effect size for short treatments ( $g+ = 0.61$ ;  $SE = .15$ ; CI from 0.32 to 0.89) was smaller than for long treatments ( $g+ = 1.43$ ;  $SE = .19$ ; CI from 1.05 to 1.80) on immediate posttests, and the confidence intervals did not overlap. A similar trend was found for the standardized-mean-gain on immediate posttests: short treatments also had smaller effects ( $g+ = 0.80$ ;  $SE = .10$ ; CI from 0.60 to 0.99) than long treatments ( $g+ = 1.72$ ;  $SE = .11$ ; CI from 1.50 to 1.94), with nonoverlapping confidence intervals, even though both effects were large. The trend was reversed somewhat for delayed posttests, where short treatments were associated with slightly smaller effects as shown in Table 18 (however, all effects were still large). The analog to ANOVA did not confirm that the defined levels of the variable corresponding to the duration of the task-based interaction treatment could account for the variability in effect sizes.

### *Effects of Other Variables*

This section addresses part of Research Question 4; in particular, it reports the results of the investigation of the potential effects of methodological and other study-related variables on study outcomes. Weighted mean effect sizes were calculated for various subsets of studies that share the same characteristics related to the publication source, language of study, country of study, basis for participant assignment to experimental versus control and comparison groups, and so forth. The effect sizes associated with different levels of these variables are presented in Table 19. (The percentages refer to the total number of studies in which the variable was represented.

Table 18

Weighted Mean Effect Sizes Associated with the Duration  
of Task-Based Interaction Treatment

Duration	Type of Effect Size	k	Effect Size		95% CI		Q <sub>W</sub>	Q <sub>B</sub>
			g <sup>+</sup>	SE	Lower	Upper		
	Between groups:							
Short	Exp – Control (Immediate)	5	0.61	.15	0.32	0.89	18.15*	11.58
Long	Exp – Control (Immediate)	5	1.43	.19	1.05	1.80	11.99	
Short	Exp – Control (Delayed)	4	1.11	.22	0.68	1.55	2.56	0.02
Long	Exp – Control (Delayed)	4	1.06	.21	0.64	1.48	8.09	
	Within groups:							
Short	Exp (Pre to Immediate)	6	0.80	.10	0.60	0.99	82.15*	37.67*
Long	Exp (Pre to Immediate)	5	1.72	.11	1.50	1.94	11.39	
Short	Exp (Pre to Delayed)	5	1.77	.15	1.49	2.06	29.71*	0.56
Long	Exp (Pre to Delayed)	5	1.63	.11	1.41	1.85	21.67*	

\* Statistically significant when the overall comparison error rate was controlled at .05 level

Some studies contributed more than one effect size because they investigated acquisition of multiple target structures.)

Based on the results of the analog to ANOVA, the only study-related variables that can explain the variability in effect sizes were student assignment to groups (random vs. nonrandom) and length of delay (short vs. long) between the instructional treatment and the delayed posttest when a delayed posttest was used in the study. The weighted mean effect size for studies utilizing nonrandom assignment of participants to experimental, control, and comparison groups  $g^+ = 1.63$  ( $SE = .19$ ) was substantially larger than the small effect  $g^+ = 0.37$  ( $SE = .10$ ) associated with random assignment. In regard to the length-of-delay variable, long-delay posttests (i.e., posttests with a delay of 28 days and over) were associated with greater weighted mean effect sizes than short-delay posttests.

The results of the analog to ANOVA were not statistically significant for the other variables; however, the face examination of the differences between the weighted mean effect sizes revealed certain trends. There were substantial differences between the

Table 19

Weighted Mean Effect Sizes Associated with Publication Type, Target Language (TL) and Language Setting, Research Setting, and Other Study-Related Variables

Variables and levels	Frequency		Effect Size		95% CI		Q <sub>W</sub>	Q <sub>B</sub>
	K	%	g <sub>+</sub>	SE	Lower	Upper		
Publication Type								0.00
Article	6	43%	0.67	.12	0.44	0.90	19.59*	
Dissertation	8	57%	0.66	.12	0.43	0.90	57.07*	
Target Language								4.76
English	4	29%	0.43	.14	0.17	0.70	15.25*	
Non-English	8	71%	0.81	.11	0.60	1.02	52.31*	
Japanese as TL								9.39
Japanese	5	36%	1.06	.16	0.76	1.36	13.59	
Non-Japanese	9	64%	0.50	.10	0.31	0.70	34.09*	
Language Context								16.11
FL	11	79%	0.89	.10	0.70	1.08	40.79*	
L2	3	21%	0.14	.16	-0.16	0.45	0.18	
Language Distance								16.99
I	5	20%	0.29	.12	0.06	0.53	16.76*	
IV	7	40%	1.05	.14	0.77	1.32	15.61	
Educational Setting								12.55
University	10	77%	0.81	.11	0.60	1.02	37.07*	
Adult Education	3	23%	0.14	.16	-0.16	0.45	0.18	
Dissertation Origin								10.75
Georgetown	5	50%	0.40	.15	0.12	0.69	19.17*	
Other	3	50%	1.27	.22	0.84	1.70	7.56	
Country								4.37
US	11	79%	0.55	.10	0.36	0.75	48.87*	
Non-US	3	21%	0.94	.16	0.64	1.24	3.83	
Assignment								34.06*
Random	8	67%	0.37	.10	0.18	0.57	14.17	
Nonrandom	4	33%	1.63	.19	1.26	2.00	4.68	
Research Setting								3.27
NS-led (Lab)	10	77%	0.74	.11	0.53	0.96	35.32*	
Learner-led	3	23%	0.41	.15	0.13	0.70	14.85*	
Length of Test Delay								27.63*
Short-Delay Tests	5	29%	0.41	.15	0.12	0.70	7.41	
Long-Delay Tests	5	36%	1.37	.22	0.93	1.80	7.43	

\* Statistically significant when the overall comparison error rate was controlled at the .05 level

insignificant (i.e., less than .20 based on Cohen's [1977] classification) effects associated with learning TL in an L2 setting  $g_+ = 0.14$  ( $SE = .16$ ; CI from -0.16 to 0.45) and the large effects associated with learning TL in an FL setting  $g_+ = 0.89$  ( $SE = .10$ ; CI from

0.70 to 1.08); and the 95% confidence intervals did not overlap. Similarly, the weighted mean effect size for adult education (e.g., ESL classes at a community center)  $g+ = 0.14$  ( $SE = .14$ ; CI from -0.16 to 0.45), which coincidentally was based on the same effect sizes as for L2, was lower considerably than for the university setting where participants were graduate and undergraduate students  $g+ = 0.81$  ( $SE = .11$ ; CI from 0.60 to 1.02). Both variables, that is, the language setting (i.e., L2 vs. FL) and the educational setting (i.e., adult education vs. university) had nonoverlapping 95% confidence intervals.

The weighted mean effect size for studies completed in the US was associated with a medium effect  $g+ = 0.43$ , whereas non-US studies yielded a large effect  $g+ = 0.94$ ; however, the latter value was based on only three effect sizes. When the effect sizes originating from the doctoral dissertations completed at Georgetown University were aggregated together, the resulting weighted mean indicated a small effect  $g+ = 0.40$  versus a large effect  $g+ = 1.27$  associated with doctoral dissertations completed at other US universities. The latter value was based on only three effect sizes, one of which was equal to 2.20 (Horibe, 2002).

The weighted mean effect size for studies that had English as the TL ( $g+ = 0.43$ ;  $SE = .14$ ; CI from 0.17 to 0.70) was smaller than for languages other than English ( $g+ = 0.81$ ;  $SE = .11$ ; CI from 0.60 to 1.02); however, the confidence intervals overlapped. Because there were five studies involving Japanese as the TL, a separate analysis was performed for these studies (see Table 19). The weighted standardized-mean-difference effect size for Japanese ( $g+ = 1.06$ ;  $SE = .16$ ; CI from 0.76 to 1.36) was greater substantially than the overall weighted mean effect size for all included studies  $g+ = 0.67$  and the weighted mean effect size for studies involving English as the TL ( $g+ = 0.43$ ). It

also was greater than the weighted mean effect size for all languages other than Japanese combined ( $g+ = 0.50$ ).

Findings regarding the language distance (i.e., the linguistic distance between the TL and the learners' L1 based on MacWhinney's [1995] classification) were consistent with these results. The included studies in which the distance between the two languages was equal to "IV" based on the classification presented in chapter III in the *Methodological Features* section ( $k = 6$ ; i.e., five studies with English learners studying Japanese and one with English learners studying Korean) had a large mean effect size for the experimental-control comparison ( $g+ = 1.05$ ;  $SE = .14$ ; CI from 0.77 to 1.32) that was greater substantially than the small mean effect size for the studies where the language distance was determined to be "I" ( $g+ = 0.29$ ;  $k = 3$ ;  $SE = .12$ ; CI from .06 to .53). The latter were the two studies involving English speakers learning Spanish (Gass & Alvarez-Torres, 2005; Toth, 2008) and Nuevo's (2006) study, in which approximately 85.00% of the learners of English were speakers of Spanish. (There were no studies with the linguistic distance of "II" and only one study with the distance of "III," specifically, Hungarian-speaking participants learning English; [Revesz, 2007].)

NS-led interaction (that typically occurs under laboratory, rather than classroom conditions) was associated with medium effects ( $g+ = 0.74$ ;  $SE = .11$ ; CI from 0.53 to 0.96) as compared with learner-led interaction (i.e., NNS learners interacting with each other in dyads or small groups) that was associated with small effects ( $g+ = 0.41$ ;  $SE = .15$ ; CI from 0.13 to 0.70). The 95% confidence intervals overlapped. The implications of these and other findings are discussed in chapter V.

Additional variables, mostly of substantive rather than methodological nature, also were listed in the Coding Form (see Appendix C) and coded when the information was available in the primary study; however, they were not included in the analysis due to insufficient aggregation. Among these variables were learner opportunity for pretask planning, presence of explicit rule review or rule modeling, learner and teacher attitudes toward task-based language teaching (TBLT), and so forth. The next section explores the potential effects of type of outcome measures used in the study on the weighted mean effect sizes associated with these specific types of outcomes.

### *Effects of Type of Outcome Measure*

This section presents data related to Research Question 5: “Is there a difference in effect size values based on what type of outcome measure (i.e., posttest measuring acquisition of the target grammatical structure) was used in the primary research study (e.g., metalinguistic judgment vs. selected response vs. oral-communication task)?” The type of outcome measure used to assess participants’ grammatical development after the treatment possibly is the most important moderator variable investigated in this meta-analysis. Therefore, presents all calculated effect sizes (between-group and within-group) in great detail, including experimental-comparison group contrasts, if the number of relevant effect sizes ( $k$ ) was equal or greater than three.

As shown in Table 20, the between-group contrasts involving both control and comparison groups were associated with small effect sizes (0.28 - 0.48) for metalinguistic-judgment tests on both immediate and delayed posttests. The 95% confidence intervals did not include zero, which means that the difference from the zero effect size was statistically significant. The within-group contrasts for the task-based



Table 20

## Weighted Mean Effect Sizes Associated with Specific Types of Outcome Measures

Type of Test	Type of Effect Size	k	Effect Size		95% CI		Q <sub>W</sub>	Q <sub>B</sub>
			g <sub>+</sub>	SE	Lower	Upper		
Between groups:								
Exp – Control (Immed.)								
Metalinguistic		8	0.36	.10	0.15	0.56	26.63*	65.95*
Free-constructed		4	2.18	.20	1.78	2.58	0.72	
Oral task		6	0.49	.13	0.24	0.74	8.87	
Exp – Comp (Immediate)								
Metalinguistic		7	0.28	.09	0.11	0.46	17.84*	29.10
Selected		3	-0.12	.24	-0.58	0.34	10.70*	
Free-constructed		3	-0.20	.21	-0.61	0.21	3.51	
Oral task		6	0.38	.16	0.07	0.69	8.67	
Exp – Control (Delayed)								
Metalinguistic		6	0.48	.14	0.20	0.75	11.75	28.71
Free-constructed		4	1.72	.22	1.30	2.15	3.54	
Oral task		6	0.38	.16	0.07	0.69	8.67	
Exp – Comp (Delayed)								
Metalinguistic		5	0.30	.10	0.09	0.50	19.81*	2.34
Free-constructed		3	-0.13	.21	-0.54	0.29	6.10	
Oral task		3	0.84	.15	0.54	1.14	2.74	
Within groups:								
Exp (Pre to Immediate)								
Metalinguistic		12	1.12	.06	0.99	1.24	131.46*	108.57*
Constrained		3	1.68	.17	1.34	2.01	3.61	
Free-constructed		5	2.70	.16	2.39	3.02	20.75*	
Oral task		7	0.88	.09	0.70	1.06	144.87*	
Exp (Pre to Delayed)								
Metalinguistic		8	0.93	.07	0.78	1.07	77.45*	91.49*
Constrained		3	1.81	.17	1.47	2.15	1.59	
Free-constructed		5	2.42	.17	2.09	2.75	27.78*	
Oral task		8	0.95	.09	0.77	1.14	130.17*	

\* Statistically significant when overall comparison error rate was controlled at .05 level

interaction (i.e., experimental) groups were large:  $g_+ = 1.12$  on immediate posttests and  $g_+ = 0.93$  on delayed posttests. There were no sufficient data to make meaningful comparisons for effect sizes associated with selected-response and constrained-constructed-response tests besides the fact that the standardized-mean-gain effect sizes for constrained-constructed response tests were large:  $g_+ = 1.68$  for immediate and  $g_+ = 1.81$  on delayed posttests.

Large weighted mean effect sizes exceeding or approximating two standard-deviation units were identified for free-constructed-response tests on comparisons between the experimental and control groups for both immediate ( $g+ = 2.18$ ) and delayed posttests ( $g+ = 1.72$ ) as well as on within-group comparisons ( $g+ = 2.70$  on immediate and  $g+ = 2.42$  on delayed posttests). The effect sizes associated with the contrast between the experimental and comparison groups, however, were negative:  $g+ = -0.20$  on immediate posttests and  $g+ = -0.13$  on delayed posttests. The 95% confidence intervals for the latter included zero, which means that these findings of negative effect sizes were not statistically trustworthy.

For the posttests that represented oral-communication tasks (and that, therefore, were most congruent with the task-based interaction treatment as discussed in chapter II under *Measures of Acquisition of Target Grammatical Structures*), the weighted standardized-mean-difference effect sizes were small (but close to medium) for the experimental-control group contrasts for immediate posttests ( $g+ = 0.49$ ) and small for delayed posttests ( $g+ = 0.38$ ). The effects were large for the experimental-comparison group contrast on immediate posttests ( $g+ = 0.85$ ) but smaller on delayed posttests ( $g+ = 0.43$ ). In line with the overall tendency, the standardized-mean-gain effect sizes for this type of outcome measure were large:  $g+ = 0.88$  for immediate posttests and  $g+ = 0.92$  for delayed posttests. The analog to ANOVA did not identify variables that could account for the variability in the effect sizes.

Based on the face examination of the data presented in Table 20, free-constructed response tests were associated with greater outcomes than oral-communication-task tests or metalinguistic-judgment tests (with the exception of the experimental-comparison

group contrasts). Oral-communication tasks had somewhat larger effects than metalinguistic-judgment tasks in regard to the standardized-mean difference, whereas metalinguistic-judgment tasks were associated with larger within-group effects (i.e., standardized mean gain) than oral-communication tasks. In view of the nonsignificant analog to ANOVA results and the fact that these findings sometimes are based on very small cell sizes (e.g.,  $k = 3$ ), they can only be interpreted as suggestive. The limitations associated with small cell sizes as well as other limitations of the present study are provided in chapter V. A brief summary of findings for all five research questions is presented in the following section.

### Summary

The search of published and unpublished literature identified 15 studies that met the inclusion and exclusion criteria for this meta-analysis that examined the effects of task-based interaction in focused oral-communication tasks used as instructional treatment to facilitate the acquisition of specific grammatical target structures in adult FL and L2 learners. In these studies, there were 22 distinct target structures for the acquisition of which associated effect sizes were calculated (standardized-mean-difference, standardized-mean-gain, or both depending upon the study design and the data reported in the study). The effect sizes calculated for individual studies were not homogeneous, which confirmed the need for a detailed analysis of potential moderator variables.

For the first question investigating the effectiveness of oral task-based interaction, the results of the meta-analysis revealed a medium effect size on immediate posttests and a large effect size on delayed posttests as compared with no instruction. The results also

showed a small effect size for task-based interaction over other types of instruction.

For the second question investigating the effectiveness of task-based interaction based on the pre- to posttest differences, task-based interaction treatment groups demonstrated substantially larger gains than control groups. The results were inconclusive regarding the comparison groups that received other types of instruction because both the task-based interaction treatment groups and comparison groups demonstrated similar, large gains on immediate and delayed posttests.

Regarding the third question investigating the potential effects of task type on the effectiveness of task-based interaction, the results showed that tasks designed on the basis of the so-called information-gap principle, including (one-way) information-gap tasks and (two-way) jigsaw tasks were associated with larger effects as compared with other types of tasks; however, the differences were not statistically significant. Closed and divergent tasks were associated with larger effects than open-ended and convergent tasks respectively (the differences between convergent and divergent tasks were statistically significant). The results of the analog to ANOVA did not confirm that any of these task-related variables could account successfully for the variability in the effect sizes.

Investigation of other potential moderator variables that was conducted for the fourth research question revealed that morphosyntactic structures tended to be associated with larger effect sizes than morphological structures, and complex structures were associated with larger effect sizes than simple structures. In most instances, the confidence intervals did not include zero and did not overlap. The results of the analog to ANOVA only confirmed that the differences between complex and simple structures could account for the variability in standardized-mean-difference effect sizes. Long-

duration treatments were associated with greater effect sizes than short treatments but only on immediate posttests, and this difference was statistically significant. The effect of the duration of the treatment tended to even out toward delayed posttests.

For the methodological variables, the analog to ANOVA confirmed that the nature of participant assignment to groups (random vs. nonrandom) and the difference between long-delay and short-delay posttests could account for the variability in the associated effect sizes. (Nonrandom assignment and long-delay posttests were associated with larger effects.) Additionally, large effects were associated with studying TL in FL settings as opposed to L2 settings and in university settings as opposed to adult education, which showed small mean effects; however, the analog to ANOVA did not confirm that the levels of these variables could account for the observed differences in effect sizes.

Finally, regarding the fifth question investigating differences in effect sizes based on what type of outcome measure was used in the primary research study, free-constructed-response measures produced larger gains in general but not on experimental-comparison group contrasts. These results are discussed in chapter V, and the limitations of the study and the implications of the findings are provided.

## CHAPTER V

### DISCUSSION, LIMITATIONS, IMPLICATIONS, AND RECOMMENDATIONS

In the introductory chapter, an argument was presented that there is insufficient evidence regarding the effectiveness of task-based interaction in form-focused instruction (FFI; see Appendix A for a list of abbreviations used in this study) of adult foreign-language (FL) and second-language (L2) learners. This meta-analysis represented an examination of experimental and quasi-experimental studies of the effectiveness of interaction that occurs in face-to-face oral-communication tasks in acquisition of specific grammatical structures of the target language (TL). This chapter includes a summary of the meta-analysis, an explanation of limitations that are likely to have influenced the results, and a discussion of the research questions with an interpretation of the results presented in chapter IV. The present chapter concludes with recommendations for future research.

#### Summary of the Meta-Analysis

An extensive review of literature located 15 empirical studies investigating the effectiveness of task-based interaction in acquisition of specific grammatical TL structures that met the criteria presented in chapter III. The inclusion and exclusion criteria were different from the previously conducted meta-analyses in the task-based interaction domain (Keck, Iberri-Shea, Tracy-Ventura, & Wa-Mbaleka, 2006; Mackey & Goo, 2007). The criteria for what language-learning activities meet the definition of an oral-communication task were more stringent (see chapter II). To generate new evidence with regard to the relationship between task-based interaction and acquisition of target

structures, both published and unpublished studies conducted between 1980 and 2009 were examined.

For the purposes of systematically collecting and analyzing data from the included primary studies, a coding form was designed (see Appendix C). The study characteristics, participant characteristics, research design features, characteristics of tasks used as the treatment, and outcome measures used to assess the participants' acquisition of the target structure were categorized, coded, and tallied across the included studies. After calculating the Hedges's  $g$  (unbiased effect size index), the findings of the eligible studies were aggregated, and potential moderator variables were analyzed. The differences in effect sizes associated with various pedagogical and methodological variables were explored using analog to analysis of variance (ANOVA).

The first research question investigated the differences between the performance of task-based interaction groups and groups that received no instruction or received other types of instruction in the target structure. The results suggest that, in line with the findings of the previous meta-analyses (Keck et al., 2006; Mackey & Goo, 2007), task-based interaction, in general, is associated with medium and large effects. As compared with other types of instruction, task-based interaction was found to be associated with small effects. The results for the second question examining within-group, pre- to posttest, gains suggest that task-based interaction treatments result in large effects as compared with small effects for groups that received no instruction. The results were inconclusive in regard to groups that received other types of instruction who also demonstrated large gains.

The results for the third, fourth and fifth questions, examining moderator

variables for possible contributions to variability, suggest that only such variables as the complexity of the target grammatical structure (complex vs. simple), the nature of participants' assignment to groups (nonrandom vs. random), and the length of delay before a delayed posttest was administered (long vs. short) could account for excess variability in the effect-size distribution.

The present chapter includes the discussion of the results presented in chapter IV for each of the five research questions. The chapter also provides a discussion of the limitations of the meta-analysis, its implications for pedagogical practice, and recommendations for further research. The following section provides a detailed presentation of the limitations.

#### Limitations of the Study

This section outlines some major limitations that are related to general methodological issues that manifested themselves in this meta-analysis as well as characteristics of the included primary studies and issues that are unique to this meta-analytic study. These limitations may have an adverse effect on the generalizability of the findings of the study.

#### *Inclusion Criteria and Search Procedures*

According to Rosenthal and DiMatteo (2001), every meta-analysis has some inherent bias by virtue of the inclusion and exclusion criteria that are set by the meta-analyst and the methods chosen to access and review the literature in the domain.

Rosenthal and DiMatteo stated that “not every computer-assisted search will be complete, and not every journal article identified” (p. 66). Inclusion criteria in the present meta-analysis were set so that only studies in which the experimental treatment verifiably



involved oral-communication tasks as defined in chapter II could be included. In some instances, primary researchers were contacted with a request for more information about the nature of the tasks.

The search of the literature was problematic in this meta-analytic study because of the lack of uniformity in terminology and study-naming conventions in the domain as well as the lack of a single definition of a task as discussed in chapter II. Most of the studies that came up in the search results based on specified key words could not be included because they investigated the effects of various types of corrective feedback that expressly was not the focus of the present meta-analysis. The independent variable in such studies typically was different types of corrective feedback while all groups of the participants received the same tasks to complete. Frequently, even the “control” group received task-based interaction but no corrective feedback. The dominance of such and other studies that were ineligible for the present meta-analysis sometimes necessitated that the meta-analyst review the full texts of dozens of studies that came up in searches without gaining a single eligible candidate. Exclusion of certain keywords, however, sometimes resulted in extremely few or no results. It is, therefore, not impossible that an eligible study that came up in a search after hundreds of ineligible ones may have been overlooked, even though the meta-analyst attempted to take every measure to safeguard against such an oversight. For example, Revesz and Hans (2006), which is a journal article, was provided by one of the co-authors who was contacted by the meta-analyst with an unrelated request, rather than located through the search process. Therefore, it is not possible to state with absolute certainty that no eligible studies have remained unidentified. The list of selected candidate studies, however, was submitted to two

renowned experts in the field who did not identify any additional studies that should be added or studies that should be removed from the list. It is hoped that this step helped safeguard the search process against possible omissions.

Even though the present meta-analysis included both published and unpublished primary research studies, as discussed under *Fail-Safe N* in chapter III, it is not protected completely from a publication bias because studies with nonsignificant results still may be less likely to be shared in any way, for example, even in the form of a conference presentation. In addition, Rosenthal and DiMatteo (2001) discussed the so-called *sophistication bias*, where research study reports that are perceived to lack sophistication may not get published or presented. This bias may provide a partial explanation for the fact that the domain is limited to very few TLs (as discussed in the *Research Synthesis*) and that laboratory, rather than classroom-based, studies carried out by seasoned researchers dominate the field.

Aside from the specified challenges, the search procedure used in this meta-analytic study cannot be considered comprehensive or exhaustive. Language acquisition research is being conducted in many parts of the world (e.g., for acquisition of English as a FL). Even though the outlined search strategy targeted some foreign publications, it is conceivable that potential candidate studies may have been published in other sources that are less known in the Western world, especially if the studies are written in languages other than English. Therefore, another limitation of the study is related to a potential retrieval bias, that is, the specified search procedure was likely to yield studies reported in English obtained primarily from sources that are well known in the Western world.

### *Small Number of Included Studies*

Because rather stringent criteria were adopted for what activities can be considered tasks, the number of located eligible studies was small ( $k = 15$ ), and the majority of the studies included in previous meta-analyses could not be included in the present study. The main reasons for a small overlap with Keck et al.'s (2006) and Mackey and Goo's (2007) meta-analyses were that these meta-analysts included studies that investigated acquisition of lexis (in addition to acquisition of grammatical target structures) and applied less stringent qualifying criteria for what activities can be considered tasks. Mackey and Goo focused exclusively on studies that investigated effectiveness of corrective feedback in task-based interaction, which was not the focus of the investigation in the present meta-analysis. Additionally, Mackey and Goo included studies of language acquisition by child (vs. adult) learners and studies that involved computer-mediated (vs. oral face-to-face) interaction.

Consequently, some of the findings were obtained based on small or very small numbers of associated effect sizes. Sometimes it was not possible to aggregate even a minimally sufficient number of studies for analysis (weighted mean effect sizes were not calculated if the number of effect sizes for a certain level of a potential moderator variable was less than three).

There were only three primary studies that involved learner-led interaction. There appears to be a larger, construct-related issue associated with a small number of studies where interaction is led by a native speaker (NS) because one of the perceived benefits of task-based language teaching (TBLT) is that it represents a learner-centered approach to instruction that requires learners to exert a greater amount of mental effort associated

with playing an active role in reaching the determined task outcome. It is doubtful that task-based interaction led by an NS always meets these expectations.

Additionally, in regard to small numbers of included studies, Norris and Ortega (2006) suggested use of caution when using inferential statistics analogous to ANOVA for moderator variables whose levels always are likely to be small. N. Ellis (2006) compared using the analog to ANOVA statistic with small numbers of studies with going on a fishing expedition and cited Keck et al. (2006) with 14 sample studies and Russell and Spada (2006) with 15 sample studies as examples. N. Ellis insisted that because the cell sizes for the  $Q$  statistic were very small, the findings should be interpreted as “usefully suggestive” but not definitive (p. 305). Dinsmore (2006), who used 17 ANOVA comparisons in his meta-analysis with 22 included samples, commented that statistically significant outcomes of repeated use of ANOVA can only be used for exploratory purposes in such cases because some outcomes may turn out to be statistically significant purely due to chance. An attempt was made to control the error rate for the comparison.

#### *Nonindependence of Study Samples and Effect Sizes*

This meta-analysis presented some challenges from the point of view of potential nonindependence of study samples. In addition to the situation where the author of an included doctoral dissertation was also one of the two co-authors of an included journal article, some of the included dissertations had the same advisor or were completed at the same university, as discussed in the *Research Synthesis* in chapter II. Rosenthal and DiMatteo (2001) recommended two ways of dealing with such issues. The first way is to “block” eligible studies originating from the same laboratory or researcher, which was not done in the present meta-analysis because it would have further reduced the number

of eligible studies. The second way is to treat the researcher or the laboratory as a potential moderator variable. Based on the small cell numbers, this analysis was not possible in some situations; however, the effect sizes originating from the three dissertations completed at Georgetown were aggregated, and the results of the comparison with other dissertations were inconclusive.

As discussed in chapter IV, effect sizes associated with different target structures generally are treated as independent in SLA meta-analyses even when they come from the same study and the same participant sample. Because some of the studies involved two or three target structures, this approach occasionally led to situations where a cell with five effect-size values only contained effect sizes from two primary research studies, which generally is not considered even minimally sufficient.

Additionally, it would be hard to argue that effect sizes for different target structures coming from the same study were unaffected completely by the characteristics unique to that study. It is more likely that there is a complex interaction between the characteristics of the target structure and various learner-, teacher-, and context-related variables. Thus, this established practice, even though it is considered to be necessary and defensible, may lead to further diminishing of the generalizability of meta-analytic findings in the domain.

Moreover, some of the moderator variables investigated in this meta-analysis potentially were not independent. For example, there was a 100% overlap between the individual effect sizes used in the calculation of the weighted mean for L2 (vs. FL) settings, on the one hand, and for adult education (i.e., one of the levels for the variable of educational setting), on the other hand, because learning a language as L2 frequently is

associated with adult education such as ESL classes at community centers. Another example of such an overlap of levels for various moderator variables is the fact that five of the 15 included studies involved Japanese as the TL. These studies constituted the majority (5 out of six) of studies that supplied effect sizes for the greatest language distance between the TL and the learners' L1 (labeled "IV").

### *Disparity of Primary Study Designs*

Research designs in the domain range from simple and straightforward to complex designs that involve numerous variables but sometimes fail to include a true control or comparison group (Keck et al., 2006; Lazaraton, 2000; Norris & Ortega, 2000). In addition to the distinction between studies with within-group and between-group designs, the present meta-analysis spanned a whole range of design features. It included studies with simple as well as very complex designs that included multiple groups, independent and dependent variables, their operationalizations, data collection procedures, outcome measures, scoring protocols, times of measurement for delayed posttests, statistical tests, and so forth. The majority of the studies investigated additional research questions that were outside the scope of the present investigation or even were unique to one particular study. Such variation can increase the generalizability of results when the data are presented clearly in the primary studies (Rosenthal & DiMatteo, 2001). In some instances, however, the meta-analyst had to sort through numerous details and go through a number of stages in the process of "shrinking" the data to a point of being able to calculate one effect size per target structure per study (as well as calculating the effect sizes for specific levels of numerous moderator variables).

The amount of within-study calculations associated with each of the studies with a

complex design sometimes was substantial and potentially may have increased the probability of error in calculations or data transfer even though the accuracy was checked routinely by the meta-analyst. The examples of specific challenges encountered by the meta-analyst in this regard are provided in the *Research Synthesis* in chapter IV. In light of the differences in operationalizations and the terminology used, establishing equivalency between constructs across the included studies for coding purposes sometimes was a challenging and painstakingly time-consuming endeavor as well.

#### *Methodological Quality of Included Studies*

For a meta-analysis, the quality of original studies can become one of its major limitations. Based on his meta-analytic examination of 30 years of interaction-based research, Plonsky (2010) reported that the findings indicated a strong relationship between study quality and study outcomes. There are different approaches to meta-analytic syntheses: some researchers (Slavin, 1986) advocated only including studies that meet strict requirements for methodological quality, and others recommended a more inclusive approach (Lipsey & Wilson, 2001). Russell and Spada (2006) explained that the reality of having small numbers of eligible studies precluded them from excluding studies on the basis of lack of adherence to methodological standards, for example, lack of prior equivalence established between groups of participants or lack of random assignment that generally are considered important quality measures (Lipsey & Wilson, 2001).

A similar issue manifested itself in the present meta-analysis. The following are some examples of potential methodological flaws found in the included studies. In at least one of the included studies, the participants in the control and experimental groups seemed to differ substantially in terms of their institutional course enrollment;

additionally, the control group did not take the delayed posttest. Another study reported the participant attrition rate of over 30% during the course of the study (which may be typical in an adult-education setting such as in ESL courses at a community center). Three of the included studies had only 6 to 8 participants in each group, and two more had 9 to 11 participants. On the opposite side of the spectrum, in one of the studies, there were 55 to 60 students in a classroom who completed the tasks in pairs or small groups. Such a large class size does not constitute a methodological flaw necessarily but the contextual variables are bound to be very different from what happens in a smaller class. Regarding the issue of using intact classes, however, even though it generally is considered to be a methodological flaw, there is an argument to be made that using intact classes contributes to the ecological validity of research (Adams, 2007) because it preserves the important contextual variables and thus makes the findings more generalizable to classroom instruction.

As discussed in the *Quality of Study* section in chapter III, Rosenthal (1991) suggested using a weighting system that takes into account methodological quality. Considering the fact that, historically, methodological quality in a strict sense is lower in SLA than in the field of cognitive psychology, no attempt to implement a weighting system was made in the present meta-analysis.

#### *High-Inference Coding Decisions*

Norris and Ortega (2000, 2006a, 2006b) pointed out the common lack of standardization and clarity in the operationalization of both independent and dependent variables in primary studies investigating effectiveness of L2 instruction. The coding decisions that required the meta-analyst and the second coder to make the greatest



number of inferences were the characteristics of tasks and the characteristics of the target structure. There were other coding decisions for which the data frequently were unclear (e.g., presence of explicit rule presentation or error correction); however, due to scarcity of data and what appeared to be little variability in these areas, these potential moderator variables were not analyzed in the present study.

For the most part, primary researchers specified, with few exceptions, the type of task or tasks they were using as the treatment based on the main design principle, for example, information-gap, jigsaw, or problem-solving tasks. This was not, however, the case with the variables of open-endedness and convergence. Very few authors provided these characteristics of the tasks or at least indicated whether the task had one acceptable outcome (closed vs. open) and whether the participants had the same assigned communicative goal in completing the task (convergent vs. divergent).

Only some of the authors specified the nature of the target structure explicitly in terms of it being morphological, syntactic, or morphosyntactic. Nevertheless, even when they did, it was evident that the primary researchers did not adhere to the same guidelines in making their determinations. In one example, the two coders disagreed with the primary researcher's designation of a structure as syntactic and labeled it morphosyntactic, which appeared to be more in line with how the structures were labeled across the studies. In some instances, the coders' designations regarding the nature of the target structure, its degree and complexity, and ambiguity were checked with the primary researchers via email but this option was not always available. The two coders' own interpretations of the features of the target structures should be treated with caution because the inferences regarding the ambiguity of the target structures were based, in

some instances, on subjective judgment and, in the case of TLs that were not known to the coders, the inferences were based exclusively on the descriptions provided by the primary researchers. Additionally, what is ambiguous to learners from a certain L1 background arguably may be less ambiguous to learners from a different L1 background.

The meta-analyst was presented with challenging decisions when determining levels of some moderator variables, for example, the levels for the duration of the treatment and for the time elapsed before the delayed posttest. Essentially, the levels for these variables were established with some consideration to the previous literature; however, because of the small number of the included studies, one of the main considerations became the need to have a minimally sufficient number of effect sizes in each cell. In view of this concern, it may be argued that the levels of the variables sometimes were set somewhat arbitrarily. Spada and Tomita (2010) acknowledged that if they had chosen a different set of criteria to distinguish between the two levels of complexity of the target structures, the results of their meta-analysis may have been different. In the present meta-analysis, the same caveat undoubtedly applies to the determination of structure complexity (that was done based on Spada and Tomita's classification) as well as to some of the other moderator variables.

### *Measurement Issues*

Due to a great variability of the posttests used in the primary studies, outcome measures were based on diverse scoring procedures and criteria. The general issues related to measuring acquisition of specific TL structures are provided in detail in chapter II under *Issues in Measuring Acquisition of Grammatical Structures*. In some of the included studies, the researchers used a stringent criterion that required the presence of at

least two instances of correct target structure use in two different posttests to show that sustained development indeed had occurred, whereas others did not employ such a criterion. Similarly, some researchers took into account both the number of attempts to produce the target structure as well as the accuracy rate in free-response or oral-communication-task posttests (e.g., Jeon, 2004), whereas others seemed to take into account only accuracy.

Chaudron (2006) pointed out that complex acquisition of linguistic items is difficult to measure in specific numerical terms because “the interaction of many components of meaning and syntactic form do not easily lend themselves to systematic quantified comparisons” (p. 326). Some of the researchers attempted to isolate various aspects of the target structure and measure improvements in each of these aspects separately. For example, in Iwashita’s (2003) study, development of locative constructions was split into development of “locative word order” and “locative particle use,” both of which were measured separately (and reported separately in Keck et al.’s meta-analysis). In other studies included in the present meta-analysis, Japanese locative constructions were treated as one structure and, for this reason, the meta-analyst in this study combined the two effect sizes by calculating their mean. Such endeavors on the part of primary researchers to investigate acquisition of specific aspects of complex structures undoubtedly were worthwhile; however, they further complicated the process of obtaining one effect size per target structure per study and were not undertaken uniformly by all primary researchers even when the target structures in question were the same or similar.

Specific types of outcome measures such as metalinguistic-judgment, selected-response, and constrained-constructed-response tests have their own inherent limitations in terms of indicating whether true acquisition has taken place, as discussed in chapter II under *Issues in Measuring Acquisition of Grammatical Structures*. A fill-in-the-blanks test may not be a reliable indicator of the learners' ability to use the associated grammatical forms correctly and appropriately in oral task-based interaction when their attention is on meaning and not on form. Therefore, the fact that more researchers in the domain use oral-communication tasks as the testing instrument is a positive development; however, as presented in chapter II, the use of tasks as the outcome measure poses its own challenges associated with task design, scoring, training of raters, feasibility, reliability, and validity.

In regard to custom-made posttests, Adams (2007), who used them in her study, pointed out that such tailored posttests may show learning on a specific item in which the learner made an error during interaction; however, they do not provide evidence that restructuring of the learner interlanguage has occurred, that is, there is no evidence that the learner is able to produce the same target structure correctly in another item, especially when the learner's primary attention is on meaning. Finally, three of the included studies used listening-comprehension tests (labeled selected-response tests because the learners had to decide whether the statements containing the target structures they heard were true to fact or not). The comprehension aspect undoubtedly is important in grammar acquisition; however, there are issues associated with assessing grammatical development through listening comprehension. Listening is defined in SLA as a very complex process that combines top-down and bottom-up processing of incoming input

and consists of three processing phases: decoding, comprehension, and interpretation (Rost, 2005), all of which present challenges to individual learners to varying degrees. Therefore, listening ability varies greatly in individual learners and there are many potentially confounding variables, especially if the learners' general listening ability is not established and taken into account prior to the treatment. Miscomprehension may occur for reasons other than lack of knowledge of grammatical structures (i.e., issues with distinguishing sounds or word boundaries); in fact, L2 learners primarily tend to use semantic (i.e., meaning-related), rather than grammatical cues to figure out meaning (Doughty & Johnston, 2006). Another example of an idiosyncratic type of outcome measure that would make it difficult to establish parallels with other measures is the unscrambling test in Silver (1999).

Reporting of the reliability and validity of outcome measures has increased considerably since Norris and Ortega's (2000) meta-analytic report; however, it was not uniform and sometimes absent in the studies included in the present meta-analysis. This and other issues related to the outcome measures in primary studies are discussed in detail in chapter II under *Issues in Measuring Acquisition of Grammatical Structures*.

#### *Missing Data for Moderator Variables*

Just as Keck et al.'s (2006), Mackey and Goo's (2007), and Russell and Spada's (2006) meta-analyses, the present meta-analysis suffered from insufficient data on identified constructs of interest. Similar to what was observed by Norris and Ortega (2000), various types of instruction were sometimes merged in a single intervention without precise control or description of its components, for example, treatments may have contained tasks preceded by, followed by, or interwoven with nontask activities of

different types. Such variations rarely were described in detail, controlled for, or systematically operationalized. Additionally, they were not aggregated sufficiently across studies to be treated as moderator variables.

As presented in the *Research Synthesis*, there were various omissions in individual studies. For example, some of the studies did not specify the basis for participant assignment to groups or the duration of the treatment. Almost no studies specified how much time within the treatment, if any, was taken by the pretask (e.g., setting up the task, learner task planning, rule review or modeling, etc.) and the posttask (e.g., reporting and receiving feedback) phases. Thus, these potentially crucial pedagogical variables could not be compared and analyzed across studies. One of the informal goals for this meta-analysis was to attempt to define the best practices of teaching grammar through interaction to the degree possible. Only three studies, however, featured real classroom contexts, and details frequently were scarce. (An example of a primary study where pedagogical features were presented in greater detail is Toth [2008]).

Learner proficiency level is another crucial variable (Porter, 1986; Williams, 1998), yet, due to lack of a uniform fine-grained classification system, in the meta-analyst's judgment it was not possible to make defensible distinctions among levels of this variable across the studies. In fact, as discussed in the *Research Synthesis* under *Learner Characteristics*, most of the studies involved beginning learners or a mixture of beginners and other levels together.

One of the coded variables in this meta-analysis (see the Coding Form in Appendix C) was task complexity that was defined in chapter II under *Cognitive*

*Complexity of the Task* as the level of information processing demands on the learners' attention, memory, and reasoning that the task imposes (Robinson, 2001a). Even though this variable in general is researched widely in SLA, the foci of the investigations typically are different somewhat from the focus of the present meta-analysis. Therefore, there were only three included studies that investigated the effects of task complexity, specifically, Kim (2009), Nuevo (2005), and Revesz (2007). (Additionally, Revesz and Han [2006] investigated effects of task content familiarity.) The individual findings in these studies provided some evidence in favor of increasing task complexity; however, there were substantial differences in operationalizations of task complexity among these studies. For this reason, and because there would not be a minimally sufficient aggregation of effect sizes for each type of effect size, no analysis for task complexity as a moderator variable could be conducted.

On a related note, as discussed under *Implications of the Study*, there was limited variety in the types of tasks used across studies. Additionally, tasks in a language curriculum frequently are chained together with each subsequent task building on the outcome of the preceding one (Nunan, 1999); however, the present meta-analysis did not address effectiveness of task sequencing due to the scarcity of data in the primary studies.

#### *Upward Bias for Standardized-Mean-Gain Effect Size*

As presented in chapter IV under *Standardized-Mean-Gain Effect Size*, the within-group, pre- to posttest effect-size measure is believed to be upwardly biased (Cheung & Chan, 2004; Gleser & Olkin, 2009; Plonsky & Oswald, forthcoming). In fact, in the present meta-analysis, unusually large effect-size values (exceeding seven standard-deviation units) were obtained on certain types of immediate oral posttests for

some of the experimental groups in Revesz and Han (2006). These results, however, had to be averaged with the effect sizes associated with the written posttest because both tests belonged to the same category in Norris and Ortega's (2000) classification and then with more tests used in the same study to obtain one effect-size value. Therefore, the resulting effect-size values actually used in the analysis were less extreme. Nevertheless, this observation of extremely large effect sizes lends credence to the concern expressed by some researchers regarding the use of standardized-mean-gain effect sizes.

This concern becomes especially relevant in view of the fact that, as explained in the section titled *Standardized-Mean-Gain Effect Size*, the appropriate formula (Lipsey & Wilson, 2001, p. 44) cannot be used to calculate the  $d$  effect-size index in the task-based interaction domain because the correlation coefficient ( $r$ ) between the pretest and the posttest scores typically is not reported in the primary studies. In the present meta-analysis, these results are provided for exploratory purposes only and are used to compare the findings of the meta-analysis with the findings of other meta-analyses where this upwardly-biased standardized-mean-gain effect size is investigated.

In conclusion, the limitations presented here and, most importantly, lack of standardization and uniformity of reporting makes comparisons of treatment effectiveness from study to study challenging and may affect adversely the reliability and validity of the present meta-analysis. Notwithstanding these limitations, it is hoped that the present meta-analytic study will contribute to broadening the scope of research into the effectiveness of form-focused task-based interaction and foster continued improvements both in pedagogical practices and the research practices in the domain. The discussion of findings by research question is provided in the following section.



## Discussion of Findings

The present meta-analysis was designed to answer five questions. In this section, the results presented in chapter IV are addressed for each of the questions.

### *Research Question 1*

*To what extent is oral task-based interaction that occurs in focused (structure-based) communication tasks (in FL and L2 instruction of adult learners) effective (i.e., how large is the standardized-mean-difference effect size resulting from task-based interaction treatments compared with other types of grammar instruction for the learners' acquisition of the target grammatical structure)?*

An examination of descriptive statistics revealed indicators that, on average, task-based interaction treatments (i.e., face-to-face focused oral-communication tasks that meet the definition for tasks presented in chapter II) result in improvement of the learners' mastery of the target grammatical structures. This finding contradicts Seedhouse's (1999; 2005) assertion that learner-to-learner interaction only leads to fossilization of faulty grammatical structures in the learner's interlanguage (i.e., the developing implicit system). Task-based interaction treatments across the included studies were associated with greater weighted mean effect sizes compared with no instruction and with other instructional activities focused on the same target structures.

When comparing interaction-treatment groups with control groups that received no instruction in the target structure, the meta-analyst obtained the weighted mean effect size  $g+ = 0.67$  (weighted mean of Hedges's [Hedges & Olkin, 1985] effect-size index  $g$  corrected for sample size), which constitutes a medium effect (Cohen, 1977). This finding is almost identical to Plonsky's (2010) finding of the average medium effect size

$d = 0.65$  (Cohen's [1977] uncorrected effect-size index  $d$ ) from the body of 174 interaction studies with diverse research purposes (both experimental and nonexperimental) published since 1980 that he examined in his meta-analysis of study quality in the task-based interaction domain.

It may seem obvious that learners who received any instruction in the target structure, including completing tasks requiring oral interaction, would learn the structure better than learners who did not receive any instruction at all. In SLA, however, the role of any formal efforts aimed at teaching grammar has been questioned, for example, by Krashen (1981, 1993) and others (e.g., Schwartz, 1993). Krashen's position that language form can only be acquired implicitly through receiving rich, comprehensible input was popular in the 1970s and 1980s but has been disputed by many SLA researchers since then. Nevertheless, in the 2000s the role of oral-communication tasks in developing grammatical competence has been questioned from the other side of the belief spectrum, that is, some researchers and practitioners hold a belief that such tasks can facilitate improvements in learners' ability to communicate but not in accuracy of grammatical form (Cobb & Lovick, 2007; 2008). Moreover, Lightbown and Spada (1993), among others, reported that some practitioners even hold a belief that task-based interaction between NNS learners may be detrimental when grammatical accuracy is the target because learners will speak ungrammatically and reinforce each others' errors. For these reasons, a medium effect size of 0.67 for experimental groups as compared with control groups is an important finding.

The associated 95% confidence interval for the 0.67 medium effect size found in the present meta-analysis was rather narrow (0.50 to 0.83), which indicates a robust

effect. These gains made by the task-based interaction groups appeared to be stable and even increased slightly, on average, on delayed posttests administered 7 to 120 days after the conclusion of the treatment. The weighted mean effect size for all delayed posttests was 0.71. More specifically, as presented in chapter IV under *Effects of Other Variables*, the weighted mean effect size for long-delayed posttests (i.e., posttests with a delay of 28 days or longer) was 1.37, which is a large effect. This finding lends support to Mackey and Goo's (2007) suggestion that, even though grammatical targets may not be associated with such large initial effects as, for example, lexical targets, these effects are durable over time.

It may be expected intuitively that task-based interaction groups would outperform control groups that received no focused instruction in the target structure. Therefore, it is important to consider the outcome of the overall effect size in regard to the comparison groups that received instruction in the same target structure but of a different kind. This overall effect size shows a small effect of 0.35 (0.47 on delayed posttests) with a relatively narrow 95% confidence interval (0.18 to 0.52); however, its lower limit is close to zero. These findings may not be considered conclusive but they suggest that interaction in specially-designed oral-communication tasks, on average, may facilitate grammar acquisition more effectively than other types of FFI, including a wide range of instructional techniques received by comparison groups. The types of instruction received by comparison groups in the present meta-analysis were mechanical drills, other traditional (i.e., nontask-based) grammar-practice activities, whole-class communicative activities, input processing, listening to unmodified or premodified input, listening to other learners' interaction and so forth. Specifically, the average effect of task-based

interaction as compared with only traditional grammar practice and mechanical drills was medium and approximately equivalent to the effect of task-based interaction over no instruction (however, it was based on only three effect sizes).

Even though task-based interaction fared quite well overall in comparison with other instructional treatments, it is not possible to state conclusively that task-based interaction was superior to all alternative types of instruction. Effect sizes for task-based interaction were negative in some individual studies for experimental-comparison contrasts. For example, in Toth's (2008) study, the participants that received whole-class, teacher-led focused interactive activities outperformed the participants that completed tasks in small learner-led groups: Hedges's  $g$  for the learner-led group was -0.66 (medium negative effect) on the metalinguistic judgment test and -0.47 (small negative effect) on the free-response test. Loschky's (1994) findings that are cited frequently to dispute the effectiveness of task-based interaction indicated a negative effect size of -0.18 (insignificant negative effect based on Cohen's [1977] classification) for the contrast with the group that received unmodified input and -0.22 (small negative effect) with the group that received premodified (i.e., elaborated, enhanced) input but no output or interaction. The group in Horibe's (2002) study that received only input outperformed the task-based interaction group with an effect size of -0.10 (insignificant negative effect based on Cohen's [1977] classification). It is important to point out, however, that in these included studies where input groups outperformed interaction groups, negative effect sizes were small or insignificant for the most part. In Toth's (2008) study where findings indicated a larger negative effect, the comparison group performed teacher-led interactive activities that essentially were very similar to small-group tasks but were conducted in a

whole-class format with what appeared to be approximately 13 or 14 participants at a time.

Keck et al. (2006) reported finding larger effects than presented in this meta-analysis based on 14 independent studies. The mean effect size for task-based interaction for grammatical target structures across all included studies was  $d = 0.94$ , where  $d$  stands for Cohen's (1977) effect-size index uncorrected for possible upward bias associated with small sample sizes. There were, however, differences in how comparisons were drawn in Keck et al.'s meta-analysis as compared with the present meta-analysis. Unlike in the present meta-analysis, Keck et al.'s reported mean effect size indicated the standardized difference in performance between task-based-interaction groups, on the one hand, and control, comparison, and even so-called baseline-interaction groups (i.e., groups that received task-based treatment that was deemed to be the least interactive among all task-based treatment groups) combined, on the other hand. Similarly to the overall effects identified in the present meta-analysis, Mackey and Goo (2007) found a medium effect for grammar ( $d = 0.59$ ;  $SD = .61$ ) on immediate posttests and larger effects on delayed posttests:  $d = 1.07$  ( $SD = .82$ ) on short-delayed (i.e., 7 to 29 days after the treatment), and  $d = 0.99$  ( $SD = .69$ ) on long-delayed posttests (i.e., 30 or more days after the treatment).

The meta-analytic findings presented in this study regarding the effectiveness of task-based interaction over other forms of instruction should not be considered definitive for the following reasons: (a) these contrasts were based on an even smaller number of effect sizes than the experimental-control contrast ( $k = 9$  for immediate posttests and  $k = 6$  for delayed posttests for comparison groups) and (b) the comparison groups themselves outperformed control groups substantially and even demonstrated greater effects (than

interaction-treatment groups) on both immediate ( $g+ = .78$ ;  $k = 6$ ) and delayed posttests ( $g+ = 1.19$ ;  $k = 5$ ).

This latter finding, to a degree, is in line with the results of Norris and Ortega's (2000) meta-analysis that indicated that, on average, explicit FFI techniques ( $d = 1.13$ ) had larger effects than implicit techniques ( $d = 0.54$ ). The 95% confidence intervals for Norris and Ortega's reported mean did not overlap, which indicated a statistically significant difference at the .05 level. In view of the fact that the protocols for most of the experimental treatments in the primary studies included in the present meta-analysis did not include explicit grammar instruction and sometimes expressly precluded it, the resulting task-based treatments most likely would fall under the implicit category. Further discussion is provided in the section titled *Implications of the Study*. To summarize, the effects of task-based interaction in focused oral-communication tasks that predisposed students to repeated use of the target structure were medium as compared with control groups and small as compared with other types of grammar treatments in this meta-analysis.

### *Research Question 2*

*Is the standardized-mean-gain effect size (i.e., effect size based on the pre- to posttest differences) larger for task-based interaction treatments as compared with other types of grammar instruction?*

The pre- to posttest gains exhibited by both task-based-interaction groups and comparison groups receiving other types of instruction in the target structures were large on both immediate ( $g+ = 1.09$  for task-based-interaction and  $g+ = 0.92$  for comparison groups) and delayed posttests ( $g+ = 1.19$  for task-based-interaction and  $g+ = 1.22$  for

comparison groups). All associated 95% confidence intervals were narrow, which indicates robust findings for both task-based-interaction and comparison groups.

Because within-group effects for task-based and comparison groups were similar, these findings suggest that all various types of FFI (both focus-on-form [FoF] and focus-on-forms [FoFS]) may have large effects on acquisition of TL grammar. This finding is congruent with Norris and Ortega's 2000 meta-analytic findings but is inconclusive with regard to the superiority of task-based interaction in comparison with other types of grammar instruction. Control groups, however, exhibited considerably smaller gains:  $g^+ = 0.16$  (insignificant effect) on immediate posttests and  $g^+ = .32$  (small effect) on delayed posttests. It is not unusual in the SLA field for control groups to show some gains in the absence of focused instruction in the target structures because all FL and L2 learners continuously receive input in the TL (which may model the use of the target structures) in the form of their teachers' and peers' talk, textbook materials, authentic materials, and so forth. Additionally, increases in the scores for all groups may occur due to the so-called test practice effect (Norris & Ortega, 2000).

The findings for the standardized-mean gains (i.e., pre- to posttest, or within-group gains) in the present meta-analysis are similar to Mackey and Goo's (2007) finding of  $d = 1.09$  and Keck et al.'s (2006) finding of  $d = 1.17$  (both indicated large effects). Neither Mackey and Goo nor Keck et al. reported mean effect sizes for control and comparison groups separately. Instead, medium effects of  $d = 0.44$  and  $d = 0.66$  were reported for the control and comparison groups together in Mackey and Goo's and Keck et al.'s meta-analyses, respectively.

When the findings of the present meta-analysis are aggregated with these previous

meta-analytic findings, there is little doubt that task-based interaction results in large within-group effects. Nevertheless, considering the inconclusive findings of the present meta-analysis in regard to experimental-comparison group contrasts and the fact that comparison groups were defined and treated differently by the previous meta-analysts, it remains unclear whether task-based interaction results in larger gains than other types of grammar instruction (even though there is some evidence in support of its potentially superior effectiveness). Suggestions for pedagogical practice are discussed under *Implications of the Study*.

### *Research Question 3*

*Is there a difference in effect-size values based on the type of focused communication task (e.g., information-gap vs. opinion-gap, closed vs. open, etc.) used in the task-based interaction treatment?*

Treatments involving tasks that were designed on the information-gap principle (i.e., information-gap tasks, jigsaw tasks, or both) were associated with greater effect-size values than the “other” types of tasks; however, this difference was not statistically significant. It is intuitive that, because of the presence of the so-called gap in what information the participants have access to, these tasks cannot be completed without one interlocutor (in one-way tasks, i.e., information-gap) or both interlocutors (in two-way tasks, i.e., jigsaw tasks) asking the other for information, thus necessarily producing TL and engaging in interaction. Therefore, these types of tasks typically are believed to push students to convey more precise information and thus lead to greater TL development (Pica, Kanagy & Falodun, 1993) than, for example, problem-solving or opinion-gap tasks



in which it possible for one of the interlocutors to contribute minimally or perhaps not at all.

These findings are not consistent with the findings by Keck et al. (2006), who reported greater effects for narrative tasks ( $d = 1.60$ ) than for jigsaw ( $d = 0.78$ ) and information-gap tasks ( $d = 0.91$ ). The findings of the present meta-analysis, however, are in line with the general belief expressed in the SLA literature that jigsaw and information-gap tasks are more beneficial to FL and L2 learners (Pica et al., 1993). Nevertheless, it is not clear whether the seeming benefits of information-gap tasks are due entirely to task design. For example, in absence of information on important learner variables (e.g., affective attitudes, cognitive readiness, and, in some cases, even learner proficiency levels), there is no certainty that well-designed problem-solving or opinion-gap tasks will not lead to similar, or even greater, benefits provided that both interlocutors are motivated and actively engaged, possess mature strategies, and purposefully strive to produce great amounts of TL. Additionally, opinion-gap tasks were not represented in the studies included in Keck et al. (2006) or in the present meta-analysis, and problem-solving tasks (based on the so-called reasoning gap) were represented minimally.

Nevertheless, consistently with Keck et al.'s (2006) findings of a greater weighted mean effect size associated with information-gap tasks (one-way) than jigsaw tasks (two-way), in the present meta-analysis, one-way tasks tended to be associated with greater gains than two-way tasks. This finding is contrary to intuitive expectations because it appears that two-way tasks would encourage both interlocutors to participate equally because both have to request and provide information. This issue still is under

investigation in SLA, and there are conflicting opinions, as presented in the section titled *One-way and Two-way Tasks* in chapter II. (For example, Gass and Varonis [1986] reported that one-way tasks had more instances of TL output being “unaccepted” by the interlocutor, which means that more elaborations were needed and more negotiation of meaning took place, whereas Long [1981] considered two-way tasks to be more beneficial based on his own empirical findings.)

The differences in effect sizes for closed versus open tasks and convergent versus divergent tasks were not investigated in the previous meta-analyses. In the present meta-analysis, closed tasks tended to be associated with somewhat greater effect sizes in most cases; however, these findings were inconclusive. SLA literature leans toward considering closed tasks that have only one possible solution to be more beneficial potentially for learners’ L2 development (Long, 1996; Loschky & Bley-Vroman, 1993), but it is conceivable that the variable of task open-endedness interacts with other variables, most importantly, learners’ TL proficiency levels (Nunan, 1991) as discussed in chapter II under *Open and Closed Tasks*.

Divergent tasks, in which participants pursued their own individually assigned goals, rather than one common goal, had greater effects in all cases, and the differences were statistically significant. A possible explanation could be that, as hypothesized by Duff (1986), divergent tasks push the interlocutors to produce more TL.

Keck et al. (2006) reported a considerable difference in effect sizes between treatments that required pushed output ( $d = 1.05$ ) and those that did not ( $d = 0.61$ ). In the present meta-analysis, pushed output was not a moderator variable because, based on the definition of oral-communication task adopted here, it was expected that the learners will

produce output and engage in interactions with their interlocutors.

#### *Research Question 4*

*Is there a difference in effect-size values based on other factors such as the type of grammatical structure targeted by the task-based-interaction treatment, duration of instruction as well as miscellaneous other teacher-related, learner-related, and contextual variables?*

It is recognized in SLA research literature that not all grammatical items in a language are “created equal,” that is, that they are diverse and unequal in terms of their learnability (DeKeyser, 1998; Doughty & Varela, 1998; R. Ellis, 2006a; VanPatten, 1996). In the present meta-analysis, morphosyntactic grammatical structures were associated with greater effects than morphological structures, and these differences were statistically significant in most cases (there were no sufficient data to include syntactic structures into this part of the analysis). A possible explanation is that morphosyntactic structures have greater perceptual saliency (Doughty & Williams, 1998) because their formation goes beyond merely adding morphemes to a word and includes sentence-level transformations, which makes them more noticeable to the learner.

Additionally, in the present meta-analysis, there was a substantial overlap between morphosyntactic structures and those labeled complex (vs. simple) based on Spada and Tomita’s (2010) classification, and complex structures were associated with statistically significantly greater effects than simple structures in this meta-analysis. Moreover, the results of the analog to ANOVA confirmed that the complexity of the structure accounted for the variability in the standardized-mean-difference effect sizes. At first glance, this finding may be contrary to intuitive expectations; however, the

learnability of a structure is affected by many interacting variables such as the learner's age, language aptitude, L1-L2 differences, and so forth (Spada & Tomita, 2010).

Complex structures typically are taught to higher-level learners who, by some accounts, are equipped better to benefit from TBLT than beginning learners (Porter, 1986; Williams, 1998), which may be one of the reasons for greater "learnability" of complex structures. Iwashita (2003) even proposed testing the so-called *threshold hypothesis* that purports that for learners to be able to benefit from interaction, they need to be at a certain threshold level of TL proficiency. Greater linguistic distance between the learners' L1 and the TL was associated with greater effect sizes. A possible explanation for this finding is that the majority of effect sizes associated with smaller linguistic distance came from studies conducted in community settings, rather than university settings, in conjunction with self-selection of university students who enroll in Category IV language courses (i.e., learners who enroll in Category IV language courses may possess higher motivation and aptitude).

No conclusions could be drawn in the present study about the moderator variable of task-essentialness of the target structure, that is, whether the use of the target structure was task-essential versus merely task-natural or task-useful. The reason was that, based on the determination of the meta-analyst and the second rater, the structure was task-essential in the majority of the studies. Moreover, only in one study did the primary researcher indicate doubt whether the participants really made use of the target structure during task completion, and many of the primary researchers had audio recordings and transcripts of the interaction in which the use of the target structure could be observed.

Long-duration task-based treatments, understandably, were associated with

greater effects than short treatments on immediate posttests, with a finding of statistical significance. On average, however, these differences appeared to decrease toward delayed posttests. The possible reasons for this decrease is that learners naturally experience backsliding after initial success with a language item (Selinker, 1972) and that learners in control groups may improve their grasp of this language item simply through exposure, even in the absence of formal instruction.

Spada and Lightbown (2008b) pointed out that FL students have fewer opportunities to use the TL outside of class than L2 students, and, therefore, FL teaching generally tends to be more form-focused than L2 teaching. In the present meta-analysis, the difference in effect sizes between the FL setting and the L2 setting was considerable and statistically significant based on nonoverlapping confidence intervals (i.e., a large effect for FL versus an insignificant effect for L2). Similarly, Mackey and Goo (2007) reported statistically significant differences favoring FL settings. These results, however, do not appear trustworthy necessarily from a practical perspective because, as pointed out in chapter IV under *Effects of Other Variables*, all included L2 studies were conducted in adult-education settings where learner-related and contextual variables undoubtedly were different from, for example, undergraduate and graduate university-education settings. Potential sources of variability between adult-education settings and university settings lies in such learner characteristics as age, language aptitude, orientation to form, personal goals and sources of motivation, and so forth. For most of these variables, the number of studies that have attempted to account for them in the domain is very small.

The mean effect size for studies conducted in laboratory settings where an NS interacted one on one with a nonnative speaker (NNS) was greater (medium effect) than

in classroom settings where NNS participants interacted with each other (small effect). This finding was not statistically significant but it is consistent with other meta-analytic evidence of larger overall effects obtained in laboratory settings (Mackey & Goo, 2007; Plonsky, 2010). Mackey and Goo pointed out that in laboratory settings, free from the distractions of the classrooms, learners may pay more attention to interactional feedback provided by the NS interlocutor. The differences between laboratory and classroom-based research and the need for more classroom-based studies are addressed further under *Limitations of the Study* in this chapter.

Contrary to the previously reported meta-analytic findings (Plonsky, 2010), studies that used intact classes rather than random assignment of participants to groups were associated with greater effect sizes in the present meta-analysis. This difference was statistically significant and the results of the analog to ANOVA confirmed that the variability in the standardized-mean-gain effect sizes can be explained by the nature of participant assignment to groups. This finding, however, was based on only four effect sizes for the nonrandom assignment level of this variable (not all the primary studies included in this meta-analysis specified the basis for group assignment). A possible explanation for this finding is that in intact classes, participants share a common context and are aware of the group dynamics that have already been established, which may affect their performance positively. Similarly to Mackey and Goo's (2007) meta-analysis in this study, conclusions could not be drawn in several important areas (e.g., for learner level as a moderator variable in this meta-analysis) due to unclear or scarce data as discussed in the section titled *Limitations of the Study*.

Finally, for the included studies that featured a delayed posttest, the length of delay appeared to play a role. Long-delay posttests were associated with larger effect sizes than short-delay posttests across the included studies. This difference was statistically significant and the analog to ANOVA confirmed that the length of delay (i.e., long vs. short delay) could account for the variability in the associated standardized-mean-difference effect sizes. This finding is consistent with Mackey and Goo's (2007) suggestion that, once a grammatical structure is acquired successfully, its mastery is not only stable but even has a tendency to improve over time.

#### *Research Question 5*

*Is there a difference in effect-size values based on what type of outcome measure (i.e., posttest measuring acquisition of the target grammatical structure) was used in the primary research study (e.g., metalinguistic judgment vs. selected response vs. oral-communication task)?*

The research synthesis conducted for the present meta-analysis (see section titled *Outcome Measures of the Research Synthesis* in chapter IV) confirmed an increase in the proportion of outcome measures congruent with communicative language teaching (CLT) that are believed to assess implicit, automatic control of processing, rather than explicit knowledge of grammatical items since Norris and Ortega completed their seminal meta-analysis in 2000. Norris and Ortega reported that discrete-point grammar tests dominated in the domain with approximately 90% of studies utilizing metalinguistic judgments (i.e., tests requiring learners to state whether a form or a sentence is grammatically correct and, in some instances, to correct the error), selected responses (e.g., choosing the correct grammatical ending or form from two or more options provided), or constrained-

constructed responses (e.g., filling in the blanks with the correct grammatical ending). A mere 10% involved extended communicative use of the TL (i.e., free-constructed responses). (In the present meta-analysis, a fifth category of outcome measure labeled “oral-communication task” was added. This category encompassed interactive face-to-face tasks that meet the criteria for tasks as defined in chapter II of this meta-analysis and are similar to the tasks used as task-based interaction treatments in the included studies.)

In the present meta-analysis, selected-response and constrained-constructed-response measures clearly were in the minority; in fact, there were insufficient numbers of primary studies utilizing each of these two types to make meaningful comparisons involving them. Metalinguistic judgment still was a popular type of test; however, only 46.67% of the included studies had a metalinguistic-judgment component. This type of test was associated with small weighted mean effect sizes for between-group experimental-control and experimental-comparison group contrasts on both immediate and delayed tests. On within-group contrasts, the effect sizes were large for both immediate and delayed meta-linguistic judgment tests.

Oral-communication tasks were used both as the treatment and as tests in 73.33% of the studies, which indicates that the domain is moving toward more communicative testing formats and that Norris and Ortega’s (2000) recommendations have been implemented by many primary researchers. This welcome trend also was pointed out by Spada and Tomita (2010) who reported that 50% of the studies included in their meta-analysis contained free-response-outcome measures. (Spada and Tomita used Norris and Ortega’s classification in its original form that did not include oral-communication tasks as a separate category).



For the most part, the outcomes for oral-communication tasks used as tests were similar to metalinguistic-judgment tests in the present meta-analysis, that is, small effect sizes were observed for between-group contrasts and large for within-group contrasts. There was only one exception to this pattern: the weighted mean effect size was large for the between-group experimental-comparison contrast on immediate posttests. This latter result could be interpreted as indicating that, although other types of instruction (i.e., input processing) undoubtedly are beneficial, they do not deliver in terms of developing learners' ability to use grammar correctly while participating in actual TL interactions as efficiently as task-based interaction treatments do. Extreme caution should be exercised, however, in making such a generalization because this large mean effect-size value was calculated based on only three contributing effect sizes.

Nevertheless, free-constructed-response measures that are next in line (after oral-communication tasks) in terms of congruence with CLT were associated with the largest overall results for between-group experimental-control comparisons and within-group comparisons (roughly around two standard-deviation units or greater). It is important to point out, however, that these results were obtained on the basis of not only oral but also written free-constructed-response tests used in the included primary studies. Perhaps this fact may help explain the negative effects sizes on the experimental-comparison contrasts. These were only a (very) small negative result on immediate posttests and an insignificant negative result on delayed posttest; however, these findings are in sharp contrast with (very) large effect sizes found for all the other types of effect sizes for free-constructed-response outcome measures (i.e., for the experimental-control contrasts and within-group contrasts).

In general, the findings regarding the effects of the specific types of outcome measures lend some support to the contention that learners who have been taught grammar communicatively, on average, can be expected to do better on communicative testing measures (Erlam, 2003; Gass & Mackey, 2007). As discussed in chapter II, the true measure of successful acquisition of a target structure is the learner's demonstrated ability for spontaneous processing of the form as it comes up in communication, rather than ability to produce the form when prompted (Jackson, 2008; Nunan, 1999). Implications for SLA theory and pedagogical practice including teacher training and curriculum development are provided in the next section.

### Implications of the Study

This section presents the theoretical and pedagogical implications of the study. The methodological implications, that is, recommendations for future research, are provided in the next section titled *Recommendations for Research*.

The present meta-analysis draws implications for the interaction hypothesis (Long, 1981; 1996) that are discussed in detail in chapter II under *Role of Interaction in Foreign and Second Language Learning*. Specifically, this study expands the empirical support for the interaction hypothesis based on both unpublished and published studies, 11 of which had not been included in any previous meta-analysis. It may be reported with caution (in view of the study's limitations presented in the previous section) that the findings lend support to the benefits of TL interaction for both FL and L2 adult learners in a variety of contexts.

The findings of the study also lend additional empirical support to the beneficial role of TBLT in general and FoF as one of its main methodological principles in

particular (Basturkmen, Loewen, & Ellis, 2004; Doughty & Williams, 1998; Long, 1991). The findings also help counteract some previously expressed concerns regarding the feasibility of communicative, task-based teaching of grammar (Seedhouse, 1999; 2005; Swan, 2005). The main concern of the opponents of TBLT typically is whether TBLT can lead to grammatical development (more so than to the development of other aspects of the TL), and this meta-analysis specifically provides some evidence that suggests that morphosyntactic (i.e., grammatical) development does occur in focused oral-communication tasks.

Even though the evidence is limited, in line with the skill acquisition theory (DeKeyser, 2007) discussed in chapter II, focused interactive tasks, more so than mechanical drills or traditional practice, appear to help learners progress to the skills of using the structure appropriately in communicative settings. As argued in chapter II, unlike traditional types of grammar practice, task-based interaction constitutes transfer-appropriate processing of TL structures that DeKeyser defined as processing that is conducive to developing skills transferrable to real-use situations in TL speaking environments.

Besides serving to reconfirm the validity of some theoretical frameworks, such as the interaction hypothesis and the FoF, the findings of the study have direct implications for pedagogical practice, specifically for the choice of activities for targeted practice of grammatical structures. As discussed in chapter I, in the obsolete, traditional view, only grammar drills and explicit discussions of grammar rules were considered to be useful for fostering learners' grammatical development (Celce-Murcia, 1992; Larsen-Freeman, 2001a; Long, 2000). In other words, whenever a need for teaching of grammar was

identified (i.e., planned teaching of a new grammatical form or intervention including an old form necessitated by learners' errors), rule explanations and traditional exercises were conducted. In chapter I, an argument was presented that TBLT can be used not only to develop the learners' fluency and general TL proficiency but also to teach specific grammatical items effectively (R. Ellis, 2003; Hinkel & Fotos, 2002; Lightbown, 2000). This assertion is supported by the aggregated findings of this meta-analysis, Keck et al.'s (2006) meta-analysis, and Mackey and Goo's (2007) meta-analysis.

Previous SLA literature pointed out the "meager" evidence of the effects of grammar instruction on learners' ability to use targeted structures in communicative-use situations (especially during unplanned use; R. Ellis, 2005). The relatively small number of free-response outcome measures identified by Norris and Ortega (2000) and the fact that they were associated with considerably smaller effects than noncommunicative outcome measures were not encouraging in this regard. The present meta-analysis has provided some, albeit limited, empirical evidence of large effects for real oral-communication tasks used as outcome measures for experimental over control and comparison groups and for free-response outcome measures for experimental over control groups. These findings suggest that teaching grammar through task-based interaction has a potential to offer substantial benefits in developing learners' ability to communicate with grammatical accuracy.

Success with grammatical development under task-based-interaction conditions, however, is not a given. It is mediated by numerous factors as shown in this and the previous meta-analyses and may be predicated on the presence of teacher's skill and experience, positive teacher and learner attitudes toward TBLT, and other factors that

could not be investigated in the present meta-analysis. The main advantage of task-based teaching of grammar is that it is a reflection of a more modern, integrated approach (FoF) that does not separate but rather unites teaching of specific grammatical forms and teaching to communicate in the TL (Doughty, 2001; Doughty & Williams, 1998; 2001; 2001; Larsen-Freeman, 2001b; 2003; Kowal & Swain, 1994; Lightbown, 2007; Long, 1996; Nassaji, 1999; Nassaji & Fotos, 2004; Nobuyoshi & Ellis, 1993; Spada, 1997; Swain & Lapkin, 2001). In this sense, simply put, task-based interaction offers language learners a double benefit. This argument is in line with Norris and Ortega's assertion that, if FoFS and FoF are equally effective in improving the mastery of the target structure (as indicated by the results of their meta-analysis), the teachers may be advised to choose FoF whenever possible because it is bound to develop learners' overall communicative competence more so than FoFS.

The findings of the present study show that learners benefit from various types of focused instruction in the target structure; however, learners who received task-based interaction had a small advantage, on average, over other learners and an even greater advantage when the effects of task-based interaction were compared specifically with the effects of mechanical drills and traditional practice.

The latter finding of larger effects of task-based interaction over traditional grammar practice is based on a small number of included studies and should not be understood to imply that any activity that presents an interactive oral-communication task is automatically beneficial for learners. In fact, it may be reasonable to assume that learners engaged in well-designed and well-implemented nontask grammar-focused activities (including input-processing and other input-based activities, whole-class

communicative activities, etc.) probably would learn more than those engaged in TBLT where task design or teaching were not carefully planned or not executed well. The same assumption may be true in situations where effectiveness of tasks is jeopardized by affective issues such as unfavorable teacher and learner attitudes.

Additionally, different types of instruction serve their own purpose, for example, input processing helps establish stronger form-meaning mappings before the learner attempts to produce the target structure (Lee & VanPatten, 2003; VanPatten, 1996). Therefore, it is not recommended that the community of practitioners juxtapose task-based grammar teaching with other types of instruction. Based on the inconclusive findings regarding the comparative effectiveness of task-based interaction versus all other types of instructional techniques present in the included studies, it may be more desirable to practice task-supported, rather than purely task-based, instruction (R. Ellis, 2003). In task-supported instruction, teachers are able to use diverse instructional elements based on their own informed decision-making, as long as, overall, the learners are primarily involved in communicating meaning rather than merely manipulating form. This suggestion is in line with Lightbown's (2007) assertion that in language learning students benefit most from being engaged in the greatest variety of types of processing in the greatest variety of contexts. As discussed in chapter II, evidence abounds in SLA literature abounds in evidence that different target structures may lend themselves to different types of instructions, and that different learners may benefit differentially from different types of instruction.

In the same vein, one specific recommendation that can be made for pedagogy is to include explicit elements of instruction into task-based activities, for example, in the

pretask or posttask phases (R. Ellis, 2003). Bearing in mind Spada and Tomita's (2010) meta-analytic findings of greater effect sizes for both simple and complex structures when they are taught explicitly as well as Norris and Ortega's (2000) findings of larger effects for explicit versus implicit instruction, it is not advisable to exclude explicit elements of instruction. This exclusion sometimes occurs based on misinterpretations of CLT as equivalent to the focus-on-meaning (FoM) approach (see *Focus on Form*, *Focus on Forms*, and *Focus on Meaning* in chapter II).

In addition to reaffirming the principles of TBLT, the present meta-analysis draws important implications regarding material development. A central concern for language teachers and curriculum developers is how to design tasks to promote learning of specific elements of the language. Some researchers argue that targeting specific grammatical structures should not be a design feature in tasks (Long, 1996; Skehan, 1998) because it detracts from the authentic communicative purpose; however, others strongly advocate such an approach (R. Ellis, 2003; Larsen-Freeman, 2001a; 2003; Nassaji, 1998). Most primary researchers whose studies were included in the meta-analysis reported that their tasks succeeded in eliciting the target structures when performed by NNS learners and frequently also by NS participants (if the task was piloted with NS interlocutors to test how likely it was to lead to the use of the target structure). There was no evidence that the approach that targets specific TL forms in so-called focused (vs. nonfocused) tasks necessarily results in artificial conversational exchanges and undermines the natural communicative purpose.

One trend that is evident from the review of the primary studies in the domain is that research is dominated by one or two types of tasks (e.g., jigsaw tasks where learners

spot the differences between two pictures). It appears that other types of tasks are underutilized and underexplored. As presented in chapter II, task-based instruction offers many design opportunities. Learners, for example, can be asked to come up with a joint plan of action, compile a ranked list of arguments, make a prediction, reach consensus on how to resolve a moral dilemma, find discrepancies between two sources of information, and so forth in the TL. The observable product of such activities can be a plan, a list, a chart, a family tree, an itinerary, a floor plan, a map, an advertisement, a description of an imaginary product, a letter, a set of instructions created by learners, a solution to a problem, an unidentified object or person, a consensus (presented verbally or in writing), and many other outcomes that are found in real-life situations outside the language classroom (R. Ellis, 2003; Leaver & Willis, 2004; Willis, 2004). It should not be assumed that designing such tasks targeting specific grammatical structures of various TLs is an easy matter; however, investigating a greater variety of tasks in research is desirable.

It also is evident that tasks still predominantly are treated in SLA as means of eliciting the learners' speech samples that then can be analyzed rather than as means of instruction. Samuda (2007) called for switching the emphasis to the pedagogical aspects of implementing tasks in real classroom contexts. It is alarming that task-based treatments in most studies are administered "cold," that is, that they do not include pretask and posttask attention to form (i.e., target structure). Bygate and Samuda (2009) warned of the danger of assuming that communication in the TL and learning of the TL are one and the same thing, whereas, in reality, learning implies a change in the learner's interlanguage that does not necessarily result from communication. Therefore, it is important to ensure that tasks are not completed for their own sake but that oral



communication, or interaction in the TL, necessarily has a learning dimension that will push the learners beyond the constraints of their current interlanguage.

Bygate and Samuda (2009) pointed out the importance of designing realistic tasks that create a pressure for learners to communicate and of equipping them with the resources they need for successful communication at the same time. Tasks may not deliver the expected learning if they fail to get the learners' interest and engagement, to be transparent in terms of the potential learning benefits, and so forth. These considerations raise the importance of the pretask phase where the teacher can set up the learners for success in various ways. Along the same lines, negotiation for meaning is not guaranteed to occur in tasks in situations where learners and teachers are bound by constraints of politeness and conscious or subconscious avoidance of the type of interaction they might perceive as being too chaotic for a classroom environment. These considerations point to importance of both teacher and learner training to ensure success in implementation of TBLT.

Regardless of the quality of task design, the classroom teacher needs to be skilled in implementation of tasks. In view of the fact that even a well-designed task in the real classroom takes on a life of its own, it is important to equip teachers with skills needed to deal with the uncertainty of learner-centered environments with unpredictable outcomes. This is a challenge for teachers who have been schooled in more traditional teaching methods (Richards, Gallo & Renandya, 2001). Teachers need rich and varied opportunities to review, experience, design, implement various classroom tasks, and reflect on their implementation (Allwright & Bailey, 1991; Bailey, 2006; Bailey, Curtis, & Nunan, 2001; Freeman & Richards, 1996; Kumaravadivelu, 1994; 2003; Larsen-

Freeman, 2001a; Richards & Lockhart, 1994). Because of the dynamic, multifaceted nature of classroom task implementation, it is not feasible to train teachers in TBLT by providing them with lockstep instructions. Teachers need to be empowered to make informed decisions as they are choosing or designing tasks targeting a specific language feature, setting up the task with the learners, priming the learners for the use of the target structure in some way if necessary, monitoring task completion and providing strategic and linguistic help without taking over, giving feedback and facilitating learner self-reflection, incorporating other instructional elements (e.g., rule modeling, input flood, traditional practice exercises, etc.), and so forth.

In terms of learner training, adult learner views on TL grammar instruction should be addressed in an ongoing dialog about how languages are learned, and learner expectations continuously should be taken into account or renegotiated when necessary because, just as teachers, learners may hold traditional views on what constitutes grammar instruction (Cobb & Lovick, 2008). Otherwise learners may circumvent the target structures, treat tasks as a diversion from “serious” learning, and so forth. The rationale for using focused tasks should be clear to the learners, which should lead to greater levels of mental and emotional investment during task completion. In general, ongoing learner strategy training that has a potential of increasing levels of deep processing and self-regulation (Vermunt & Vermetten, 2004) should be an integral part of a language course.

To summarize, in the words of N. Ellis (2006) who commented on Keck et al.’s (2006) meta-analytic results, it has been demonstrated that conscious learning in social interactions that serve to scaffold learner comprehension and production promote the

acquisition of the target structures. Nevertheless, the complexity of task-based teaching of grammar should not be underestimated. It is easy to take an ideological position but difficult to understand the impact of various factors and their multifaceted interactions with each other. There is no set of well-designed tasks and no prescriptions that could guarantee success with improving learners' grammatical accuracy. Effective teachers typically subscribe to an eclectic approach to grammar teaching where they draw on a variety of instructional techniques depending on the goals of the program as well as the needs, cognitive styles, and inclinations of individual students (Purpura, 2004). The presence of a great number of diverse and dynamically interacting factors that mediate the success of grammar acquisition rule out the possibility of prescriptive lockstep procedures for teaching TL grammar. Moreover, excessive reliance on one approach or strict administrative exclusions of certain methodologies may not be productive ultimately because of the diverse characteristics of target structures and learners. Integration of a variety of creative techniques offers a greater potential for empowering teachers to make online decisions about meeting diverse and evolving learner needs, as long as the guiding principle of CLT, such as teaching the language through communication as much as possible, is followed.

This section has addressed the implications of the present meta-analytic study for practitioners, faculty trainers, and curriculum developers. The next section presents recommendations for future research in the field of task-based interaction in SLA.

#### Recommendations for Research

Based on the findings and implications of this study, as well as on the careful analysis of its apparent limitations, recommendations for future research are provided in

this section. The data presented in the *Research Synthesis* completed for this meta-analysis (see chapter IV) have provided evidence that research practices in the domain have improved somewhat in accordance with the recommendations made in the seminal meta-analytic study completed by Norris and Ortega (2000). Meta-analysts have reported that interaction researchers now pay greater attention to reporting means and standard deviations,  $t$  values, and exact  $p$  values (that can be used to calculate effect sizes) as well as to issues of reliability and validity (Mackey & Gass, 2006; Plonsky, 2010; Russell & Spada, 2006). Thus, as pointed out by Plonsky (2010), the “meta-analyzability” of primary interaction-based research is increasing. Nevertheless, some flaws undoubtedly remain, and the most important recommendations for further improvement are listed below.

More experimental and quasi-experimental research studies are needed in the task-based interaction domain. Plonsky (2010) pointed out that interaction as an area of research has exhibited a distinct preference for nonexperimental research and relied heavily on observational or ex post facto designs. Out of the 174 studies included in Plonsky’s (2010) meta-analysis, only 66 (38%) were experimental studies.

A related issue is that a very limited number of TLs were represented in the present meta-analysis, Keck et al.’s (2006) meta-analysis, and Mackey and Goo’s (2007) meta-analysis. The only three TLs represented in Keck et al.’s study were English (as FL and L2), Spanish, and Japanese. In addition to these languages, Mackey and Goo had one study involving French and the present meta-analysis had one study involving Korean. Clearly, this is a very small and nonrepresentative cross-section of the world languages, and primary studies are needed involving other Asian and European languages, Russian

and other Slavic languages, Arabic, Persian-Farsi, and so forth.

Statistical reporting should be improved further. Chapelle and Duff (2003) provided detailed requirements for manuscripts submitted for publication in *TESOL Quarterly* in which authors were instructed to report, among other results, the power and the effect sizes resulting from all statistical tests. Ideally, primary researchers should report both uncorrected effect-size values (Cohen's  $d$ ) and corrected, unbiased (Hedges's  $g$ ) values. Whenever possible, it also is desirable to report pretest-posttest correlation coefficients ( $r$ ), which would allow for the proper effect-size formula to be applied when calculating standardized-mean gain effect sizes (Lipsey & Wilson, 2001, p. 44). Additionally, standard error of the mean and confidence intervals also should be reported (Mackey & Goo, 2007).

If the research domain moves toward greater uniformity of research designs and improved statistical reporting, it may be possible to implement Rosenthal's (1991) recommendation to apply a weighting system when coding for methodological quality in future meta-analyses. If a greater number of experimental and quasi-experimental research studies investigating the effects of task-based TL interaction on acquisition of specific grammatical structures are conducted, it may be possible to follow the strict guidelines for reporting the mean effect sizes only for sets of values that are found to be homogeneous based on the test of homogeneity (i.e.,  $Q$  statistic; Lipsey & Wilson, 2001).

Chaudron (2006) pointed out the growing consistency in descriptive classifications of the factors involved in language acquisition as a welcome development. Nevertheless, there is still much work to be done for the research domain to agree upon consistent empirical operationalizations of its central constructs so that variables can be

replicated across learner populations, contexts, and so forth (Norris & Ortega, 2000). For example, a more precise definition of learner proficiency levels (e.g., beginning, low intermediate, high intermediate, advanced, etc.) may allow researchers to test Iwashita's (2003) threshold hypothesis that purports that learners need to be at a certain threshold level of TL proficiency in order to benefit fully from task-based instruction.

For the purposes of further improving the reporting conventions, researchers should specify such information as the type of treatment task from the point of view of as many relevant classification systems as possible. As presented under *Limitations of the Study* in this chapter, lack of the descriptors provided by the primary researchers led to the need to make many high-inference decisions during the coding process. It also would be helpful if primary researchers specified certain characteristics of target structures according to uniform classification systems, for example, the one for structure complexity proposed by Spada and Tomita (2010) to assist readers and meta-analysts who are not familiar with the TL of the study. Including such descriptors would minimize guesswork, especially when detailed descriptive information about the treatment tasks is scarce due to space limitations or other reasons.

In general, future studies seeking to investigate the effectiveness of task-based interaction need to increase the level of detail and report as many potential moderator variables as possible, including the origin of the task, teacher and learner familiarity with TBLT and attitudes toward it, learner cognitive characteristics (e.g., aptitude, field-independence, working memory), and so forth. According to Mackey and Polio (2009), these are factors that can affect learners' access to feedback, input, and output and can cause them to pay more or less attention to features in the input. Other under-researched

variables hypothesized to affect interaction are learner gender (Gass & Varonis, 1985; Pica, Holliday, Lewis, & Morgenthaler, 1989), age (Han, 2004), L1-L2 differences (Seol, 2007), pair groupings (in terms of the interlocutor's TL level; Kim, 2009), and so forth. In other words, the findings presented in this study can be considered baseline information, and other researchers can build on these results. Future research should focus on complex, multifaceted aspects of interaction, which is difficult if the research domain remains scarce.

Regarding the outcome measures used in primary research, wide variability in the types of tests used may account for variability in results. Therefore, researchers should attempt to streamline the testing measures as much as possible. This is not an easy task because proficiency tests with established reliability and validity (e.g., TOEFL) are a poor measure of acquisition of individual structures. As a minimum, the researchers should report reliability information and adhere to the classification of outcome measures originally offered by Norris and Ortega (2000) and used by Keck et al. (2006) and in the present meta-analysis. Oral-communication tasks (added to Norris and Ortega's classification in the present meta-analysis) that represent an authentic outcome measure most congruent with task-based interaction treatments should be used as much as possible in addition to free-constructed-response measures in Norris and Ortega's classification that may be more limited in scope and noninteractive by nature. It is possible that more standardized measures will be developed at least for widely researched English structures (e.g., questions, past tense, locative prepositions).

To test Mackey and Goo's (2007) assertion that interaction effects are delayed but durable for grammar (as opposed to, for example, lexis) that received some evidence-

based support in the present meta-analysis, a more systematic, uniform planning of long-delayed posttests would be desirable to include tests with a delay of 60 days or more.

Finally, future research needs to adopt a stronger connection with pedagogy. To this end, classroom-based studies where the teacher is the only proficient TL speaker and where interaction occurs in NNS-NNS dyads or groups are needed if the research goal is to understand the nature and effects of interaction in the classroom (Spada & Lightbown, 2008b). Based on her own findings that negotiation for meaning did not occur among NNS learners, Foster (1998) questioned whether findings obtained in laboratory settings could be applied to classroom contexts. Additionally, some of the researchers who conducted laboratory-based studies expressed doubt whether their treatments could be replicated in classroom settings (Mackey & Goo, 2007). In laboratory studies, NS interlocutors are in charge and follow strict protocols, including executing instructions that would not make pedagogical sense in a classroom (e.g., not providing feedback on learner errors in the target structure or switching topics in case of conversation breakdowns). Investigating task-based interaction in short sessions that do not include a pretask or posttask phase helps control the variables but represents a poor reflection of real classroom teaching and does not take into account important teacher-, learner-, and context-related characteristics. Plonsky (2010) asserted that an increase in classroom-based research indicates a domain's theoretical maturity; therefore, a welcome development would be an increase in the numbers of classroom-based studies as opposed to laboratory studies.

### Conclusion

The contention in this meta-analysis was that task-based interaction as an



instructional technique is beneficial not only for developing the learners' overall proficiency in the TL but also for facilitating the development of learners' mastery of specific grammatical structures when specially-designed, high-quality focused tasks are used. This contention is supported by evidence in the present study, especially when this evidence is aggregated with the findings from the previous meta-analyses in the task-based interaction domain.

The findings in the present meta-analysis prohibited a firm declaration that task-based interaction was more effective than other instructional techniques to be made simply on the basis of this study. The meta-analytic findings were interpreted as suggestive that instruction that integrates many diverse techniques may be beneficial for development of FL and L2 grammatical competence as long as development of learners' communicative competence is not neglected or short-changed. It was further suggested that teachers and curriculum developers should include explicit focus-on-form into task-based language teaching in the form of integrated, rather than isolated, grammar teaching (Spada & Lightbown, 2008a). Future research should not focus primarily on seeking to investigate effectiveness of task-based interaction in focused oral-communication tasks as compared with other types of instruction but mostly on examining what factors contribute to effectiveness of task-based interaction in teaching grammar. Fellow researchers are encouraged to contribute to defining potential moderator variables to allow for aggregation of greater numbers of studies with clearly defined levels of these variables for subsequent meta-analyses.

## REFERENCES

\*References marked with an asterisk indicate studies included in the meta-analysis.

Adair-Hauck, B., & Donato, D. (2002). The PACE Model: A story-based approach to meaning and form for standards-based language learning. *The French Review*, 76(1), 265-276.

\*Adams, R. (2007). Do second language learners benefit from interaction with each other? In A. Mackey (Ed.), *Conversational interaction in second language acquisition* (pp. 29-52). Oxford, UK: Oxford University Press.

Allwright, R., Bailey, K.M. (1991). *Focus on the language classroom: An introduction to classroom research for language teachers*. Cambridge, UK: Cambridge University Press.

Anderson, J. (1983). *The architecture of cognition*. Cambridge, MA: Harvard University Press.

Anderson, J. (1993). *Rules of the mind*. Hillsdale, NJ: Lawrence Erlbaum.

Anderson, L.W., & Krathwohl, D. R. (2001). *A taxonomy for learning, teaching and assessing: A revision of Bloom's taxonomy of educational objectives*. New York: Longman.

Aston, G. (1986). Trouble-shooting in interaction with learners: The more the merrier? *Applied Linguistics*, 7(2), 128-143. doi:10.1093/applin/7.2.128

Ayoun, D. (2001). The role of negative and positive feedback in the second language acquisition of passé composé and imparfait. *The Modern Language Journal*, 88(1), 31-55. doi:10.1111/0026-7902.00106

Bailey, K. M. (1996). Working for washback: A review of the washback concept in language testing. *Language Testing*, 13(3), 257-279. doi:10.1177/026553229601300303

Bailey, K. M. (2006). *Language teacher supervision: A case-based approach*. Cambridge, UK: Cambridge University Press.

Bailey, K. M., Curtis, A., & Nunan, D. (2001) *Pursuing professional development: The self as source*. Boston: Heinle & Heinle.

Basturkmen H., Loewen, S., & Ellis, R. (2004). Teachers' stated beliefs about incidental focus on form and their classroom practices. *Applied Linguistics*, 25(2), 243-272. doi:10.1093/applin/25.2.243

- Berben, M., Van den Branden, K., & Van Gorp, K. (2007). "We'll see what happens." Tasks on paper and tasks in a multilingual classroom. In K. Van den Branden, K. Van Gorp, & M. Verhelst (Eds.), *Tasks in Action. Task-based language education from a classroom-based perspective* (pp. 32-67). Cambridge, UK: Cambridge Scholars Publishing.
- Bialystock, E. (1988). Psycholinguistic dimensions of second language proficiency. In W. Rutherford & M. Sharwood-Smith (Eds.), *Grammar and Second Language Teaching* (pp. 31-50). New York: Newbury House.
- Bialystock, E. (1994a). Analysis and control in the development of second language proficiency. *Studies in Second Language Acquisition*, 16(2), 157-168.
- Bialystock, E. (1994b). Representation and ways of knowing: Three issues in second language acquisition. In N. C. Ellis (Ed.), *Implicit and explicit learning of languages* (pp. 549-569). London: Academic Press.
- Bloom, B. S. (1956). *Taxonomy of educational objectives: Cognitive domain*. New York: Longman.
- Bloomfield L. (1961). *Language*. New York: Holt, Rinehart & Winston.
- Brandl, K. (2008). *Communicative language teaching in action: Putting principles to work*. Upper Saddle River, NJ: Pearson Education.
- Breen, M. (1987). Learner contributions to task design. In C. Candlin & D. Murphy (Eds.), *Language learning tasks* (pp. 23-46). Englewood Cliffs, NJ: Prentice Hall.
- Breen, M. (1989). The evaluation cycle for language learning tasks. In R. K. Johnson (Ed.), *The second language curriculum* (pp. 187-206). Cambridge, UK: Cambridge University Press.
- Breen, M. P., & Candlin, C. N. (1980). The essentials of a communicative curriculum in language teaching. *Applied Linguistics*, 1(2), 89-112. doi:10.1093/applin/1.2.89
- Brown, H. D. (2001). *Teaching by principles: An interactive approach to language pedagogy* (2<sup>nd</sup> ed.). New York: Longman.
- Brown, J. D., & Hudson, T. (1998). The alternatives in language assessment. *TESOL Quarterly*, 32(4), 653-675. doi:10.2307/3587999
- Brumfit, C. (1984). *Communicative methodology in language teaching: The role of fluency and accuracy*. Cambridge, UK: Cambridge University Press.
- Bruton, A., & Samuda, V. (1980). Learner and teacher roles in the treatment of oral errors in group work. *RELC Journal*, 11(2), 49-63.

- Byrd, P. (1998). Grammar in the foreign language classroom: Making principled choices. Center for Applied Linguistics. Retrieved March 13, 2008, from <http://www.nclrc.org/essentials/index.htm>.
- Bygate, M., & Samuda, V. (2009). Creating pressure in task pedagogy: The joint roles of field, purpose, and engagement within the interaction approach. In A. Mackey & C. Polio (Eds.), *Multiple perspectives on interaction: Second language research in honor of Susan M. Gass* (pp. 90-116). New York: Routledge.
- Byrd, P. (2005). Instructed grammar. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 545-561). Mahwah, NJ: Lawrence Erlbaum.
- Cadierno, T. (1995). Formal instruction from a processing perspective: An investigation into the Spanish past tense. *The Modern Language Journal*, 79(2), 179-193. doi:10.2307/329618
- Cameron, J., & Epling, W. F. (1989). Successful problem solving as a function of interaction style for non-native students of English. *Applied Linguistics*, 11(4), 392-406. doi:10.1093/applin/10.4.392
- Canale, M., & Swain, M. (1980). Theoretical basis of communicative approaches. *Applied Linguistics*, 1(1), 1-47.
- Carroll, J. B. (1990). Cognitive abilities in foreign language aptitude: Then and now. In T. S. Parry & C. W. Stansfield (Eds.), *Language aptitude reconsidered* (pp. 11-29). Englewood Cliffs, NJ: Prentice Hall.
- Carroll, S., & Swain, M. (1993). Explicit and implicit negative feedback: An empirical study of the learning of linguistic generalizations. *Studies in Second Language Acquisition*, 15(3), 357-386. doi:10.1017/S0272263100012158
- Celce-Murcia, M. (1992). Under what circumstances, if any, should formal grammar instruction take place? Formal grammar instruction: An educator's comments. *TESOL Quarterly*, 26(2), 406-408.
- Chapelle, C. A., & Duff, P. A. (2003). Some guidelines for conducting quantitative and qualitative research in TESOL. *TESOL Quarterly*, 37, 157-178.
- Chaudron, C. (1985). A method for examining the input/intake distinction. In S. M. Gass & C. G. Madden (Eds.), *Input in second language acquisition* (pp. 285-302). Rowley, MA: Newbury House.
- Chaudron, C. (2003). Data collection in SLA research. In C. Doughty & M. Long (Eds.), *The handbook of second language acquisition* (pp. 762-828). Malden, MA: Blackwell.

- Chaudron, C. (2006). Some reflections on the development of (meta-analytic) synthesis in second language research. In J. Norris & L. Ortega (Eds.), *Synthesizing research on language learning and teaching* (pp. 323-340). Philadelphia: John Benjamins.
- Cheung, S. F., & Chan, D. K. (2004). Dependent effect sizes in meta-analysis: Incorporating the degree of interdependence. *Journal of Applied Psychology*, 89, 780-791.
- Csikszentmihalyi, M. (1996). *Creativity: Flow and the psychology of discovery and invention*. New York: HarperCollins.
- Cobb, M. (2004). Input elaboration in second and foreign language teaching. *Dialog on Language Instruction*, 16(1), 13-23.
- Cobb, M., & Lovick, N. (2007). The concept of foreign language task: Misconceptions and benefits in implementing task-based instruction. *Bridges*, 21, 7-14.
- Cobb, M., & Lovick, N. (2008). Current issues in teaching of grammar: Some pedagogical possibilities for integrated form-focused instruction. *Bridges*, 22, 1-15.
- Cohen, J. (1977). *Statistical power analysis for the behavioral sciences*. New York: Academic Press.
- Cooper, H. (1998). *Synthesizing research: A guide for literature reviews* (3<sup>rd</sup> ed.). Thousand Oaks, CA: Sage.
- Cooper, H. (2003). Editorial. *Psychological Bulletin*, 129, 3-9.
- Coughlan, P., & Duff, P. (1994). Same task, different activities: Analysis of SLA from an activity theory perspective. In J. Lantolf & G. Appel (Eds.), *Vygotskian approaches to second language research* (pp. 173-194). Norwood, NJ: Ablex.
- Craik, F. (2002). Levels of processing: Past, present... and future? In M. Conway (Ed.), *Levels of processing 30 years on* (pp. 305-318). Hove, East Essex, UK: Psychology Press.
- Craik, F., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General*, 104, 268-294.
- Crookes, G. (1989). Planning and interlanguage variability. *Studies in Second Language Acquisition*, 11, 367-383.
- Curtiss, S. (1988). Abnormal language acquisition and the modularity of language. In F. Newmeyer (Ed.), *Linguistics: The Cambridge survey. Linguistic theory: Extensions and implications*, Vol. 2 (pp. 96-116). Cambridge, UK: Cambridge

University Press.

Cuskelly, E., & Gregor, S. (1994). Perspectives on computer-mediated communication. In T. Evans & D. Murphy (Eds.), *Research in distance education* (pp. 115-126). Geelong, Victoria, Australia: Deakin University Press.

de la Fuente, M. J. (2002). Negotiation and oral acquisition of L2 vocabulary: The roles of input and output in the receptive and productive acquisition of words. *Studies in Second Language Acquisition*, 24(1), 81-112.  
doi:10.1017/S0272263102001043

DeKeyser, R. M., & Sokalski, K. J. (1996). The differential role of comprehension and production practice. *Language Learning*, 46(4), 613-642. doi:10.1111/j.1467-1770.1996.tb01354.x

DeKeyser, R. M. (1997). Beyond explicit rule learning: Automatizing second language morphosyntax. *Studies in Second Language Acquisition*, 19(2), 195-221.  
doi:10.1017/S0272263197002040

DeKeyser, R. M. (1998). Beyond focus on form: Cognitive perspectives on learning and practicing in L2 grammar. In C. Doughty & J. Williams (Eds.), *Focus on form in classroom second language acquisition* (pp. 42-63). Cambridge, UK: Cambridge University Press.

DeKeyser, R. M. (2000). The robustness of critical period effects in second language acquisition. *Studies in Second Language Acquisition*, 22(4), 493-533.

DeKeyser, R. M. (2001). Automaticity and automatization. In P. Robinson (Ed.), *Cognition and Second Language Instruction* (pp. 125-151). Cambridge, MA: Cambridge University Press.

DeKeyser, R. M. (2003). Implicit and explicit learning. In C. J. Doughty & M. H. Long (Eds.), *The handbook of second language acquisition* (pp. 313-348). Malden, MA: Blackwell.

DeKeyser, R. M. (2005). What makes learning second-language grammar difficult? A review of issues. *Language Learning*, 55, *Supplement 1*, 1-25.  
doi:10.1111/j.0023-8333.2005.00294.x

DeKeyser, R. M. (2007). Situating the concept of practice. In R. DeKeyser (Ed.), *Practice in a second language: Perspectives from applied linguistics and cognitive psychology* (pp. 1-18). Cambridge, UK: Cambridge University Press.

De Ridder, I., Vangehuchten, L., & Gomez, M. S. (2007). Enhancing automaticity through task-based language learning. *Applied Linguistics*, 28(2), 309-315.  
doi:10.1093/applin/aml057

- Dinsmore, T. (2006). Principles, parameters, and SLA: A retrospective meta-analytic investigation into adult L2 learners' access to Universal Grammar. In J. Norris & L. Ortega (Eds.), *Synthesizing research on language learning and teaching* (pp. 53-90). Philadelphia: John Benjamins.
- Dornyei, Z. (2002). The motivational basis of language learning tasks. In P. Robinson (Ed.), *Individual differences and instructed language learning* (pp. 137-158). Amsterdam: John Benjamins Publishing Company.
- Doughty, C. J. (2001). Cognitive underpinnings of focus on form. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 206-257). Cambridge, UK: Cambridge University Press.
- Doughty, C. J., & Long, M. H. (2001). Optimal psycholinguistic environments for distance foreign language learning. *University of Hawaii Working Papers in ESL*, 20. Retrieved March 20, 2004, from [http://www.hawaii.edu/sls/uhwpsel/20\(1\)/Doughty&Long.doc](http://www.hawaii.edu/sls/uhwpsel/20(1)/Doughty&Long.doc)
- Doughty, C. J., & Long, M. H. (Eds.). (2006). *The handbook of second language acquisition*. Malden, MA: Blackwell Publishing.
- Doughty, C., & Pica, T. (1986). "Information gap" tasks: Do they facilitate second language acquisition? *TESOL Quarterly*, 20(2), 305-325. doi:10.2307/3586546
- Doughty, C., & Varela, E. (1998). Communicative focus on form. In C. Doughty & J. Williams (Eds.), *Focus on form in classroom second language acquisition* (pp. 114-138). Cambridge, UK: Cambridge University Press.
- Doughty, C., & Williams, J. (Eds.). (1998). *Focus on form in classroom second language acquisition*. Cambridge, UK: Cambridge University Press.
- Duff, P. (1986). Another look at interlanguage talk: Taking task to task. In R. Day (Ed.), *Talking to learn: Conversation in second language acquisition* (pp. 147-181). Rowley, MA: Newbury House.
- Egi, T. (2007). Recasts, learners' interpretations and L2 development. In A. Mackey (Ed.), *Conversational interaction in second language acquisition* (pp. 249-268). Oxford, UK: Oxford University Press.
- Elder, C., Erlam, R., & Philp, J. (2007). Explicit language knowledge and focus on form: Options and obstacles for TESOL teacher trainees. In S. Fotos & H. Nassaji (Eds.), *Form-focused instruction and teacher education. Studies in honor of Rod Ellis* (pp. 225-240). Oxford, UK: Oxford University Press.
- Ellis, N. (1995). Consciousness in second language acquisition: A review of field studies and laboratory experiments. *Language Awareness*, 4(3), 123-146.

- Ellis, N. (2006). Meta-analysis, human cognition, and language learning. In J. Norris & L. Ortega (Eds.), *Synthesizing research on language learning and teaching* (pp. 301-322). Philadelphia: John Benjamins.
- Ellis, R. (1989). Are classroom and naturalistic language acquisition the same? A study of the classroom acquisition of German word order rules. *Studies in Second Language Acquisition*, 11(3), 305-328.
- Ellis, R. (1994a). *The study of second language acquisition*. Oxford, UK: Oxford University Press.
- Ellis, R. (1994b). A theory of instructed second language acquisition. In N. C. Ellis (Ed.), *Implicit and explicit learning of languages*. London: Academic Press.
- Ellis, R. (2001). *Form-focused instruction and second language learning*. Malden, MA: Blackwell Publishers.
- Ellis, R. (2002). The place of grammar instruction in the second/foreign language curriculum. In E. Hinkel & S. Fotos (Eds.), *New perspectives on grammar teaching in second language acquisition in second language classrooms* (pp. 17-34). Mahwah, NJ: Lawrence Erlbaum.
- Ellis, R. (2003). *Task-based language learning and teaching*. New York: Oxford University Press.
- Ellis, R. (2005). Measuring implicit and explicit knowledge of a second language: A psychometric study. *Studies in Second Language Acquisition*, 27(2), 141-172. doi:10.1017/S0272263105050096
- Ellis, R. (2006a). Current issues in the teaching of grammar: An SLA perspective. *TESOL Quarterly*, 40(1), 83-107.
- Ellis, R. (2006b). Modelling learning difficulty and second language proficiency: The differential contributions of implicit and explicit knowledge. *Applied Linguistics*, 27(3), 431-463. doi:10.1093/applin/aml022
- Ellis, R. (2006c). Researching the effects of form-focussed instruction on L2 acquisition. *AILA Review*, 19(1), 18-41.
- Ellis, R. (2007). The differential effects of corrective feedback on two grammatical structures. In A. Mackey (Ed.), *Conversational interaction in second language acquisition* (pp. 339-360). Oxford, UK: Oxford University Press.
- Ellis, R., Loewen, S., & Erlam, R. (2006). Implicit and explicit corrective feedback and the acquisition of L2 grammar. *Studies in Second Language Acquisition*, 28(2), 339-368. doi:10.1017/S0272263106060141



- Erlam, R. (2003). Evaluating the relative effectiveness of structured-input and output-based instruction in foreign language learning. *Studies in Second Language Acquisition*, 25(4), 559-582.
- Foster, P. (1998). A classroom perspective on the negotiation of meaning. *Applied Linguistics*, 19(1), 1-23. doi:10.1093/applin/19.1.1
- Foster, P., & Skehan, P. (1996). The influence of planning and task type on second language performance. *Studies in Second Language Acquisition*, 18, 229-323.
- Fotos, S. (1994). Integrating grammar instruction and communicative language use through grammar consciousness-raising tasks. *TESOL Quarterly*, 28(2), 323-351.
- Fotos, S. (2002). Structure-based interactive tasks for the EFL grammar learner. In E. Hinkel & S. Fotos (Eds.), *New perspectives on grammar teaching in second language acquisition in second language classrooms* (pp. 135-154). Mahwah, NJ: Lawrence Erlbaum.
- Fotos, S., & Ellis, R. (1991). Communicating about grammar: A task-based approach. *TESOL Quarterly*, 25(4), 605-628.
- Freeman, D. A., & Richards, J. C. (Eds.) (1996). *Teacher learning in language teaching*. Cambridge, UK: Cambridge University Press.
- Fujii, A. (2005). *Individual differences in task performance: Aptitude profiles, orientation to form, and second language production in the EFL classroom* (Doctoral dissertation). Retrieved from Dissertations & Theses: A&I. (AAT 3230980)
- Garcia, P., & Asencion, Y. (2001). Interlanguage development of Spanish learners: Comprehension, production, and interaction. *Canadian Modern Language Review*, 57(3), 377-402.
- Gass, S. M. (1988). Integrating research areas: A framework for second language studies. *Applied Linguistics*, 9(2), 198-217.
- Gass, S. M. (1997). *Input, interaction and the second language learner*. Mahwah, NJ: Lawrence Erlbaum.
- Gass, S. M., & Lewis, K. (2007). Perceptions of interactional feedback: Differences between heritage language learners and non-heritage language learners. In A. Mackey (Ed.), *Conversational interaction in second language acquisition: A collection of empirical studies*. Oxford, UK: Oxford University Press.
- Gass, S., & Mackey, A. (2007). *Data elicitation for second and foreign language research*. Mahwah, NJ: Lawrence Erlbaum.

- Gass, S., & Selinker, L. (2008). *Second language acquisition: An introductory course*. New York: Routledge.
- \*Gass, S., & Alvarez-Torres, M. (2005). An investigation of the ordering effect of input and interaction. *Studies in Second Language Acquisition*, 27(1), 1-31.  
doi:10.1017/S0272263105050011
- Gass, S., & Varonis, E. M. (1985). Task variation and non-native/non-native negotiation of meaning. In S. Gass & C. Madden (Eds.), *Input in second language acquisition* (pp. 149-161). Rowley, MA: Newbury House.
- Gass, S., & Varonis, E. M. (1989). Incorporated repairs in nonnative discourse. In M. Eisenstein (Ed.), *The dynamic interlanguage* (pp. 71-86). New York: Plenum.
- Gass, S., & Varonis, E. M. (1994). Input, interaction, and second language production. *Studies in Second Language Acquisition*, 16(3), 283-302.
- Gleser, L. J., & Olkin, I. (2009). Stochastically dependent effect sizes. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2<sup>nd</sup> ed., pp. 357-376). New York: Sage.
- Greenberg, J. (Ed.). (1978). *Universals of human language* (Vols. 1-4). Stanford, CA: Stanford University Press.
- Han, Z. (2004). *Fossilization in adult second language acquisition*. Clevedon, UK: Multilingual Matters.
- Harley, B. (1986). *Age in second language acquisition*. Clevedon, UK: Multilingual Matters.
- Hedges, L. V. (1981). Distribution theory for Glass' estimator of effect size and related estimators. *Journal of Educational Statistics*, 6(2), 107-128.
- Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic Press.
- Higgs, T. V., Clifford, R. (1982). The push toward communication. In T.V. Higgs (Ed.) *Curriculum, competence, and the foreign language teacher* (pp. 57-79). Lincolnwood, IL: National Textbook Company.
- Hinkel, E., & Fotos, S. (2002). From theory to practice: A teacher's view. In E. Hinkel & S. Fotos (Eds.), *New perspectives on grammar teaching in second language acquisition in second language classrooms* (pp. 1-12). Mahwah, NJ: Lawrence Erlbaum.

- \*Horibe, S. (2002). *The output hypothesis and cognitive processes: An examination via acquisition of Japanese temporal subordinate conjunctions* (Doctoral dissertation). Retrieved from Dissertations & Theses: A&I. (AAT 3104957)
- Howatt, A. (1984). *A history of English language teaching*. Oxford, UK: Oxford University Press.
- Hulstijn, J., & de Graaff, R. (1994). Under what conditions does explicit knowledge of a second language facilitate the acquisition of implicit knowledge? A research proposal. *AILA Review*, 11(1), 97-112.
- Hyltenstam, K., & Abrahamsson, N. (2006). Maturational constraints in SLA. In C. J. Doughty & M. H. Long (Eds.), *The handbook of second language acquisition* (pp. 539-588). Malden, MA: Blackwell Publishing.
- Iwashita, N. (2003). Negative feedback and positive evidence in task-based interaction: Differential effects on L2 development. *Studies in Second Language Acquisition*, 25(1), 1-36. doi:10.1017/S0272263103000019
- Jackson, C. (2008). Proficiency level and the interaction of lexical and morphosyntactic information during L2 sentence processing. *Language Learning*, 58(4), 875-909. doi:10.1111/j.1467-9922.2008.00481.x
- Jeon, E. H., & Kaya, T. (2006). Effects of L2 instruction on interlanguage pragmatic development: A meta-analysis. In J. Norris & L. Ortega (Eds.), *Synthesizing research on language learning and teaching* (pp. 165-212). Philadelphia: John Benjamins.
- \*Jeon, K. (2004). *Interaction-driven learning: Characterizing linguistic development* (Doctoral dissertation). Retrieved from Dissertations & Theses: A&I. (AAT 3137061)
- Johnson, R. L., Penny, J. A., & Gordon, B. (2009). *Assessing performance: Designing, scoring, and validating performance tasks*. New York: The Guilford Press.
- Jourdenais, R., Ota, M., Stauffer, S., Boyson, B., & Doughty, C. (1995). Does textual enhancement promote noticing? A think aloud protocol analysis. In R. Schmidt (Ed.), *Attention and awareness in foreign language learning* (pp. 183-216). Honolulu, HI: University of Hawaii Press.
- Keck, C., Iberri-Shea, G., Tracy-Ventura, N., & Wa-Mbaleka, S. (2006). Investigating the empirical link between task-based interaction and acquisition: A meta-analysis. In J. Norris & L. Ortega (Eds.), *Synthesizing research on language learning and teaching* (pp. 91-131). Philadelphia: John Benjamins.
- Kellerman, E. (1985). If at first you do succeed. In S. Gass & C. Madden (Eds.), *Input in second language acquisition* (pp. 345-353). Rowley, MA: Newbury House.

- Kempe, V., & Brooks, P. J. (2008). Second language learning of complex inflectional systems. *Language Learning*, 58(4), 703-746. doi:10.1111/j.1467-9922.2008.00477.x
- Kim, J-H., & Han, Z. (2007). Recasts in communicative EFL classes: Do teacher intent and learner interpretation overlap? In A. Mackey (Ed.), *Conversational interaction in second language acquisition* (pp. 269-297). Oxford, UK: Oxford University Press.
- \*Kim, Y. (2009). *The role of task complexity and pair grouping on the occurrence of learning opportunities and L2 development* (Doctoral dissertation). Retrieved from Dissertations & Theses: A&I. (AAT 3370628)
- Kowal, M., & Swain, M. (1994). Using collaborative language production tasks to promote students' language awareness. *Language Awareness*, 3(2), 73-93.
- \*Koyanagi, K. (1998). *The effects of focus-on-form tasks on the acquisition of a Japanese conditional "to": Input, output and "task-essentialness"* (Doctoral dissertation). Retrieved from Dissertations & Theses: A&I. (AAT 9920550)
- Krashen, S. (1981). *Second language acquisition and second language learning*. Oxford, UK: Pergamon.
- Krashen, S. (1982). *Principles and practice in second language acquisition*. Oxford, UK: Pergamon Press.
- Krashen, S. (1985). *The input hypothesis: Issues and implications*. Oxford, UK: Longman.
- Krashen, S. (1993). The effect of formal grammar teaching: Still peripheral. *TESOL Quarterly*, 26(2), 409-411.
- Krouglov, A., & Kurylko, K. (1999). Linguistics without tears: Incorporating theory into practice. In J. Davie, N. Landsman, & L. Silvester (Eds.), *Russian language teaching methodology and course design* (pp. 31-39). Nottingham, UK: Astra Press.
- Kruschke, J. (2005). Category learning. In K. Lamberts & R. Goldstone (Eds.), *Handbook of cognition* (pp. 183-201). London: Sage Publications.
- Kumaravadelu, B. (1994). The postmethod condition: Emerging strategies for second/foreign language teaching. *TESOL Quarterly*, 28(1), 27-48.
- Kumaravadelu, B. (2003). *Beyond methods: Microstrategies for language teaching*. New Haven, CT: Yale University Press.

- Kwon, E.-Y. (2005). The “natural order” of morpheme acquisition: A Historical survey and discussion of three putative determinants. *Working Papers in TESOL & Applied Linguistics*, 5(1), 1-21.
- Lane, S., & Stone, C.A. (2006). Performance assessments. In R. L. Brennan (Ed.), *Educational measurement* (4<sup>th</sup> ed., pp. 387-431). Westport, CT: American Council on Education & Praeger.
- Larsen-Freeman, D. (1995). On the teaching and learning of grammar: Challenging the myths. In F. R. Eckman (Ed.), *Second language acquisition theory and pedagogy* (pp. 131-150). Mahwah, NJ: Lawrence Erlbaum.
- Larsen-Freeman, D. (2001a). “Grammarizing” in the ESL Classroom. Archived webcast. Retrieved November 10, 2001, from the Heinle & Heinle Publishers website <http://www.connectlive.com/events/heinle/registered.html>.
- Larsen-Freeman, D. (2001b). Teaching grammar. In M. Celce-Murcia (Ed.), *Teaching English as a second or foreign language* (3rd ed., pp. 251-266). Boston: Heinle & Heinle.
- Larsen-Freeman, D. (2003). *Teaching language: From grammar to grammarizing*. Boston: Heinle & Heinle.
- Larsen-Freeman, D., & Long, M. (1991). *An introduction to second language acquisition*. London: Longman.
- Lazaraton, A. (2000). Current trends in research methodology and statistics in applied linguistics. *TESOL Quarterly*, 34, 175-181.
- Leaver, B. (2000). Cognitive and affective issues in the learning and teaching of Slavic languages: A response. In B. Rifkin, O. Kagan, & S. Bauckus (Eds.), *The learning and teaching of Slavic languages and cultures* (pp. 215-228). Bloomington, IN: Slavica.
- Leaver, B. L., & Kaplan, M. A. (2004). Task-based instruction in U.S. Government Slavic language programs. In B. L. Leaver & J. R. Willis (Eds.), *Task-based instruction in foreign language education* (pp. 47-66). Washington, DC: Georgetown University Press.
- Leaver B. L., & Willis, J. R. (Eds.). (2004). *Task-based instruction in foreign language education*. Washington, DC: Georgetown University Press.
- Lee, J. (2000). *Tasks and communicating in language classrooms*. Boston: McGraw-Hill.
- Lee, J., & VanPatten, B. (2003). *Making communicative language teaching happen*. New York: McGraw Hill.

- Leeman, J. (2007). Feedback in L2 learning: Responding to errors during practice. In R. DeKeyser (Ed.), *Practice in a second language: Perspectives from applied linguistics and cognitive psychology* (pp. 111-137). Cambridge, UK: Cambridge University Press.
- Lightbown, P. M. (1983). Exploring relations between developmental and instructional sequences in L2 acquisition. In H. G. Seliger and M. H. Long (Eds.), *Classroom-oriented research in second language acquisition* (pp. 217-243). Rowley, MA: Newbury House.
- Lightbown, P. M. (2000). Anniversary article: Classroom research and second language teaching. *Applied Linguistics*, 21(4), 431-462. doi:10.1093/applin/21.4.431
- Lightbown, P. (2007, February). *Putting form-focused instruction in its proper place*. Plenary session conducted at the Defense Language Institute Foreign Language Center, Presidio of Monterey, CA.
- Lightbown, P., & Spada, N. (1993). *How languages are learned*. Oxford, UK: Oxford University Press.
- Lipsey, M.W., & Wilson, D.B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.
- Littlewood, W. (2004). The task-based approach: Some questions and suggestions. *ELT Journal*, 58(4), 319-326.
- Long, M. H. (1981). Input, interaction, and second language acquisition. In H. Winitz (Ed.), *Native language and foreign language acquisition* (pp. 259-278). New York: New York Academy of Sciences.
- Long, M. (1982). Native speaker/non-native speaker conversation in the second language classroom. In M. Long & J. Richards (Eds.), *Methodology in TESOL: A book of readings* (pp. 339-354). New York: Newbury House.
- Long, M. H. (1985). A role for instruction in second language acquisition: Task-based language teaching. In K. Hyltenstam & M. Pienemann (Eds.), *Modeling and assessing second language acquisition* (pp. 77-99). Clevedon, UK: Multilingual Matters.
- Long, M. H. (1989). Task, group, and task-group interactions. *University of Hawaii Working Papers in ESL*, 8(2), 1-26.
- Long, M. H. (1990). Maturational constraints on language development. *Studies in Second Language Acquisition*, 12, 251-286.

- Long, M. H. (1991). Focus on form: A design feature in language teaching methodology. In K. DeBot, R. Ginsberg, & C. Kramsch (Eds.), *Foreign language research in cross-cultural perspective* (pp. 39-52). Philadelphia: John Benjamins.
- Long, M. H. (1993). Second language acquisition as a function of age: Research findings and methodological issues. In K. Hyldenstam & A. Viberg (Eds.), *Progression and regression in language* (pp. 196-221). Cambridge, UK: Cambridge University Press.
- Long, M. H. (1996). The role of the linguistic environment in second language acquisition. In W. Ritchie & T. Bhatia (Eds.), *Handbook of second language acquisition* (pp. 413-468). San Diego, CA: Academic Press.
- Long, M. H. (1997). Focus on form in Task-Based Language Teaching. *Fourth Annual McGraw-Hill Satellite Teleconference*. Retrieved October 15, 2006, from <http://www.mhhe.com/socscience/foreignlang/top.htm>.
- Long, M. H. (2000). Focus on form in Task-Based Language Teaching. In R. D. Lambert & E. Shohamy (Eds.), *Language policy and pedagogy. Essays in honor of A. Ronald Walton* (pp. 179-192). Philadelphia: John Benjamins.
- Long, M. H. (2006). Stabilization and fossilization in second language development. In C. J. Doughty & M. H. Long (Eds.), *The handbook of second language acquisition* (pp. 487-536). Malden, MA: Blackwell Publishing.
- Long, M. H. (2007). *Problems in SLA*. Mahwah, NJ: Lawrence Erlbaum.
- Long, M. H., & Crookes, G. (1993). Units of analysis design: The case for task. In G. Crookes & S. Gass (Eds.), *Tasks in a pedagogical context: Integrating theory and practice* (pp. 9-54). Clevedon, UK: Multilingual Matters.
- Long, M. H., Inagaki, S., & Ortega, L. (1998). The role of implicit negative feedback in SLA: Models and recasts in Japanese and Spanish. *The Modern Language Journal*, 82(3), 357-371. doi:10.2307/329961
- Long, M. H., & Norris, J. (2000). Task-based teaching and assessment. In M. Byram (Ed.), *Encyclopedia of language teaching* (pp. 597-603). London: Routledge.
- Long, M. H., & Robinson, P. (1998). Focus on form: Theory, research, and practice. In C. Doughty & J. Williams (Eds.), *Focus on form in classroom second language acquisition* (pp. 15-41). Cambridge, UK: Cambridge University Press.
- \*Loschky, L. (1994). Comprehensible input and second language acquisition: What is the relationship? *Studies in Second Language Acquisition*, 16(3), 303-325.

- Loschky, L., & Bley-Vroman, R. (1993). Grammar and task-based methodology. In G. Crookes & S. Gass (Eds.), *Tasks and language learning* (pp. 123-167). Clevedon, UK: Multilingual Matters.
- Lyster, R., & Mori, H. (2006). Interactional feedback and instructional counterbalance. *Studies in Second Language Acquisition*, 28(2), 269-300. doi:10.1017/S0272263106060128
- Lyster, R., & Ranta, L. (1997). Corrective feedback and learner up-take: Negotiation of form in communicative classrooms. *Studies in Second Language Acquisition*, 19(1), 37-61.
- Macaro, E. (2003). Theories, grammar and methods. In E. Macaro (Ed.), *Teaching and learning a second language: A guide to recent research and its applications* (pp. 21-61). London: Continuum.
- \*Mackey, A. (1999). Input, interaction and second language development. *Studies in Second Language Acquisition*, 21(4), 557-587.
- Mackey, A. (2002). Beyond production: Learners' perceptions about interactional processes. *International Journal of Educational Research*, 37(3), 379-394. doi:10.1016/S0883-0355(03)00011-9
- Mackey, A. (2006). Feedback, noticing and second language development: An empirical study of L2 classroom interaction. *Applied Linguistics*, 27(3), 405-430. doi:10.1093/applin/ami051
- Mackey, A. (2007). *Conversational interaction in second language acquisition*. Oxford, UK: Oxford University Press.
- Mackey, A., & Gass, S. M. (2005). *Second language research: Methodology and design*. Mahwah, NJ: Lawrence Erlbaum.
- Mackey, A., & Gass, S. M. (2006). Pushing the methodological boundaries in interaction research: An introduction to the special issue. *Studies in Second Language Acquisition*, 28(2), 169-178.
- Mackey, A., & Goo, J. (2007). Interaction research in SLA: A meta-analysis and research synthesis. In A. Mackey (Ed.), *Conversational interaction and second language acquisition. A series of empirical studies* (pp. 377-419). New York: Oxford University Press.
- Mackey, A., & Oliver, R. (2002). Interactional feedback and children's L2 development. *System*, 30(4), 459-477. doi:10.1016/S0346-251X(02)00049-0
- Mackey, A., Oliver, R., & Leeman, J. (2003). Interactional input and the incorporation of feedback: An exploration of NS-NNS and NNS-NNS adult and child dyads.



*Language Learning*, 53(1), 35-66. doi:10.1111/1467-9922.00210

- Mackey, A., & Philp, J. (1998). Conversational interaction and second language development: Recasts, responses, and red herrings? *The Modern Language Journal*, 82(3), 338-356.
- Mackey, A., & Polio, C. (2009). Introduction. In A. Mackey & C. Polio (Eds.), *Multiple perspectives on interaction: Second language research in honor of Susan M. Gass* (pp. 1-10). New York: Routledge.
- MacWhinney, B. (1995). Language specific prediction in foreign language learning. *Language Testing*, 12(3), 292-320.
- McDonough, K. (2006). Interaction and syntactic priming: English L2 speakers' production of Dative constructions. *Studies in Second Language Acquisition*, 28(2), 179-207. doi:10.1017/S0272263106060098
- McLaughlin, B. (1987). *Theories of second language learning*. London: Edward Arnold.
- Mitchell, R., & Myles, F. (1998). *Second language learning theories*. London: Edward Arnold.
- Montgomery, C., & Eisenstein, M. (1985). Real reality revisited: An experimental communicative course in ESL. *TESOL Quarterly*, 19(2), 317-334.
- Morris, S. B. (2008). Estimating effect sizes from pretest-posttest-control group designs. *Organizational Research Methods*, 11, 364-386.
- Nagata, N. (1993). Intelligent computer feedback for second language instruction. *The Modern Language Journal*, 77(3), 330-339.
- Nagata, N. (1998). Input vs. output practice in educational software for second language acquisition. *Language Learning and Technology*, 1(2), 23-40.
- Nakahama, Y., Tyler, A., & van Lier, L. (2001). Negotiation of meaning in conversational and information gap activities: A comparative discourse analysis. *TESOL Quarterly*, 35(3), 377-405.
- Nassaji, H. (1999). Towards integrating form-focussed instruction and communicative interaction in the second language classroom: Some pedagogical possibilities. *Canadian Modern Language Review*, 55(3), 385-402.
- Nassaji, H., & Fotos, S. (2004). Current developments in the teaching of grammar. *Annual Review of Applied Linguistics*, 24, 126-145. doi:10.1017/S0267190504000066

- Newell, A., & Rosenbloom, P. (1981). Mechanisms of skill acquisition and the law of practice. In R. J. Anderson (Ed.), *Cognitive skills and their acquisition* (pp. 1-55). Hillsdale, NJ: Lawrence Erlbaum.
- Newport, E. (1990). Maturational constraints on language learning. *Cognitive Science*, 14(1), 11-28.
- Nobuyoshi, J., & Ellis, R. (1993). Focused communication tasks and second language acquisition. *ELT Journal*, 47(3), 203-210.
- Norris, J., & Ortega, L. (2000). Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis. *Language Learning*, 50(3), 417-528.  
doi:10.1111/0023-8333.00136
- Norris, J., & Ortega, L. (2006a). Defining and measuring SLA. In C. J. Doughty & M. H. Long (Eds.), *The handbook of second language acquisition* (pp. 717-761). Malden, MA: Blackwell Publishing.
- Norris, J., & Ortega, L. (Eds.). (2006b). *Synthesizing research on language learning and teaching*. Philadelphia: John Benjamins.
- \*Nuevo, A. M. (2006). *Task complexity and interaction: L2 learning opportunities and development* (Doctoral dissertation). Retrieved from Dissertations & Theses: A&I. (AAT 3247335)
- Nunan, D. (1989). *Designing tasks for the communicative classroom*. Cambridge, UK: Cambridge University Press.
- Nunan, D. (1991). Communicative tasks and the language curriculum. *TESOL Quarterly*, 25(2), 279-295.
- Nunan, D. (1993). Task-based syllabus design: Selecting, grading and sequencing tasks. In G. Crooks, & S. Gass (Eds.), *Tasks in a pedagogical context: Integrating theory and practice* (pp. 55-68). Clevedon, UK: Multilingual Matters.
- Nunan, D. (1999). *Second language teaching and learning*. Boston: Heinle & Heinle.
- Nunan, D. (2004). *Task-based language teaching*. Cambridge, UK: Cambridge University Press.
- Nunan, D. (2006). Task-based language teaching in the Asia context: Defining 'task'. *Asian EFL Journal*, 8 (3). Retrieved September 16, 2007, from [http://www.asian-efl-journal.com/Sept\\_06\\_dn.php](http://www.asian-efl-journal.com/Sept_06_dn.php).
- O'Rourke, B. (2005). Form-focused interaction in online tandem learning. *CALICO Journal*, 22(3), 433-466.

- Parker, K., & Chaudron, C. (1987). The effects of linguistic simplifications and elaborative modifications on L2 comprehension. *University of Hawaii Working Papers in ESL*, 6, 107-133.
- Pavesi, M. (1986). Markedness, discursial modes, and relative clause formation in a formal and an informal context. *Studies in Second Language Acquisition*, 8(1), 38-55.
- Pica, T. (1991). Classroom interaction, participation and comprehension: Redefining relationships. *System*, 19(4), 437-452.
- Pica, T. (1994). Research on negotiation: What does it reveal about second-language learning conditions, processes, and outcomes? *Language Learning*, 44(3), 493-527.
- Pica, T. (2009, June). *Form focusing tasks: Their multiple roles and contributions to input comprehension, output production, and language learning outcomes*. Plenary session conducted at the Defense Language Institute Foreign Language Center, Presidio of Monterey, CA.
- Pica, T., & Doughty, C. (1988). Variations in classroom interaction as a function of participant pattern and task. In J. Fine (Ed.), *Second language discourse* (pp. 41-55). Norwood, NJ: Ablex.
- Pica, T., Holliday, L., Lewis, N., Berducci, D., & Newman, J. (1991). Language learning through interaction: What role does gender play? *Studies in Second Language Acquisition*, 13(3), 343-376.
- Pica, T., Holliday, L., Lewis, N., & Morgenthaler, L. (1989). Comprehensible input as an outcome of linguistic demands on the learner. *Studies in Second Language Acquisition*, 11(1), 63-90.
- Pica, T., Kanagy, R., & Falodun, J. (1993). Choosing and using communication tasks for second language instruction. In G. Crookes & S. M. Gass (Eds.), *Tasks and language learning* (pp. 9-34). Clevedon, UK: Multilingual Matters.
- Pica, T., Kang, H.-S., & Sauro, S. (2006). Information gap tasks: Their multiple roles and contributions to interaction research methodology. *Studies in Second Language Acquisition*, 28(2), 301-338. doi:10.1017/S027226310606013X
- Pica T., Young R., & Doughty, C. (1987). The impact of interaction on comprehension. *TESOL Quarterly*, 21(4), 737-758.
- Pienemann, M. (1984). Psychological constraints on the teachability of languages. *Studies in Second Language Acquisition*, 6(2), 186-214.

- Pienemann, M. (1989). Is language teachable? *Applied Linguistics*, 10(1), 52-79.
- Pienemann, M., & Johnston, M. (1987). Factors influencing the development of language proficiency. In D. Nunan (Ed.), *Applying second language acquisition research* (pp. 45-141). Adelaide, Australia: National Curriculum Resource Center, AMEP.
- Plonsky, L. (2010). *30 years of interaction: Research methods, study quality, and outcomes*. Unpublished manuscript, Michigan State University, East Lansing, MI.
- Plonsky, L., & Oswald, F. L. (forthcoming). How to do a meta-analysis. In A. Mackey & S. M. Gass (Eds.), *A guide to research methods in second language acquisition*. London: Basil Blackwell.
- Porter, P. (1986). How learners talk to each other: Input and interactions in task-centered discussions. In R. R. Day (Ed.), *Talking to learn: Conversation in second language acquisition* (pp. 200-224). Rowley, MA: Newbury House.
- Prabhu, N. S. (1987). *Second Language Pedagogy*. Oxford, UK: Oxford University Press.
- Purpura, J. E. (2004). *Assessing grammar*. Cambridge, UK: Cambridge University Press.
- \*Revesz, A. (2007). Focus on form in task-based language teaching: Recasts, task complexity, and L2 learning (Doctoral dissertation). Retrieved from Dissertations & Theses: A&I. (AAT 3287867)
- \*Revesz, A., & Han, Z-H. (2006). Task content familiarity, task type and efficacy of recasts. *Language Awareness*, 15, 160-179.
- Richards, J. C., Gallo, P. B., & Renandya, W. A. (2001). Exploring teachers' beliefs and the processes of change. *The PAC Journal* 1(1), 1-41.
- Richards, J. C., & Lockhart, C. (1994). *Reflective teaching in second language classroom*. Cambridge, UK: Cambridge University Press.
- Robinson, P. (1996). Learning simple and complex second language rules under implicit, incidental, enhanced, and instructed conditions. *Studies in Second Language Acquisition*, 19(2), 223-247.
- Robinson, P. (2001a). Task complexity, cognitive resources, and syllabus design: A triadic framework for investigating task influences on SLA. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 287-318). New York: Cambridge University Press.
- Robinson, P. (2001b). Task complexity, task difficulty, and task production: Exploring interactions in a componential framework. *Applied Linguistics*, 22(1), 27-57. doi:10.1093/applin/22.1.27

- Robinson, P. (2005). Cognitive abilities, chunk-strength, and frequency effects in implicit artificial grammar and incidental L2 learning: Replications of Reber, Walkenfeld, and Hernstadt (1991) and Knowlton and Squire (1996) and their relevance for SLA. *Studies in Second Language Acquisition*, 27(2), 235-268.  
doi:10.1017/S0272263105050126
- Robinson, P. (2007). Criteria for classifying and sequencing pedagogic tasks. In M. del Pilar Garcia Mayo (Ed.), *Investigating tasks in formal language learning* (pp. 7-27). Clevedon, UK: Multilingual Matters.
- Rodimkina, A. (1999). Syntactic interference in learning Russian by English-speaking students. In J. Davie, N. Landsman, & L. Silvester (Eds.), *Russian language teaching methodology and course design* (pp. 41-53). Nottingham, England: Astra Press.
- Rosa, E., & O'Neill, M. D. (1999). Explicitness, intake, and the issue of awareness: Another piece to the puzzle. *Studies in Second Language Acquisition*, 21(4), 511-556.
- Rosenthal, R. (1991). *Meta-analytic procedures for social research* (Rev. ed.). Thousand Oaks, CA: Sage.
- Rosenthal, R., & DiMatteo, M.R. (2001). Meta-analysis: Recent developments in quantitative methods for literature reviews. *Annual Review of Psychology*, 52, 59-82.
- Rost, M. (2005). L2 Listening. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 503-527). Mahwah, NJ: Lawrence Erlbaum.
- Russell, J., & Spada, N. (2006). The effectiveness of corrective feedback for second language acquisition: A meta-analysis of the research. In J. M. Norris & L. Ortega (Eds.), *Synthesizing research on language learning and teaching* (pp. 133-164). Philadelphia: John Benjamins.
- Rutherford, W. (1987). *Second language grammar learning and teaching*. New York: Longman.
- Salaberry, M.R. (1997). The role of input and output practice in second language acquisition. *The Canadian Modern Language Review*, 53(2), 422-451.
- Samuda, V. (2005). Expertise in pedagogic task design. In K. Johnson (Ed.), *Expertise in second language learning and teaching* (pp. 230-254). Oxford, UK: Oxford University Press.

- Samuda, V. (2007, September). *Tasks, design and the architecture of pedagogic spaces*. Plenary session conducted at the 2<sup>nd</sup> International TBLT Conference. University of Hawaii.
- Samuda, V., Gass, S., & Rounds, P. (1996, March). *Two types of task in communicative language teaching*. Paper presented at the TESOL convention, Chicago.
- Savignon, S. (1972). *Communicative competence: An experiment in foreign language teaching*. Philadelphia: Center for Curriculum Development.
- Savignon, S. (1983). *Communicative competence: Theory and classroom practice*. Reading, MA: Addison-Wesley.
- Savignon, S. (2001). Communicative language teaching for the twenty-first century. In M. Celce-Murcia (Ed.), *Teaching English as a second or foreign language* (3rd ed., pp. 13-28). Boston: Heinle & Heinle.
- Schmidt, R.W. (1983). Interaction, acculturation, and the acquisition of communicative competence: A case study of an adult. In N. Wolfson & E. Judd (Eds.), *Sociolinguistics and second language acquisition* (pp. 137-174). Rowley, MA: Newbury House.
- Schmidt, R. (1990). The role of consciousness in second language learning. *Applied Linguistics*, 11(2), 129-158.
- Schmidt, R. (1993). Awareness and second language acquisition. *Annual Review of Applied Linguistics*, 13, 206-226.
- Schmidt, R., & Frota, S. (1986). Developing basic conversational ability in a second language: A case study of an adult learner of Portuguese. In R. Day (Ed.), *Talking to learn: conversation in second language acquisition* (pp. 237-326). Rowley, MA: Newbury House.
- Schumann, J. (1979). The acquisition of English negation by speakers of Spanish: A review of the literature. In R. W. Andersen (Ed.), *The acquisition and use of Spanish and English as first and second languages* (pp. 3-32). Washington, DC: TESOL.
- Schwartz, B. (1993). On explicit and negative data effecting and affecting competence and linguistic behavior. *Studies in Second Language Acquisition*, 15, 147-163.
- Seedhouse, P. (1999). Task-based interaction. *ELT Journal*, 53(3), 149-156.
- Seedhouse, P. (2005). "Task" as research construct. *Language Learning*, 55(3), 533-570. doi:10.1111/j.0023-8333.2005.00314.x

- Segalowitz, N. (2003). Automaticity and second languages. In C. J. Doughty & M. H. Long (Eds.), *The handbook of second language acquisition* (pp. 382-408). Malden, MA: Blackwell.
- Selinker, L. (1972). Interlanguage. *International Review of Applied Linguistics*, 10(3), 209-230.
- Seol, H. (2007). *The impact of age and L1 influence on L2 ultimate attainment* (Doctoral dissertation). Retrieved from Dissertations & Theses: A&I. (AAT 3259258)
- Sharwood-Smith, M. (1981). Consciousness raising and the second language learner. *Applied Linguistics*, 2, 159-168. doi:10.1093/applin/2.2.159
- Sharwood-Smith, M. (1988). Consciousness raising and the second language learner. In W. Rutherford & M. A. Sharwood-Smith (Eds.), *Grammar and second language teaching: A book of readings* (pp. 51-60). New York: Newbury House.
- Sharwood-Smith, M. (1993). Input enhancement in instructed SLA: Theoretical bases. *Studies in Second Language Acquisition*, 15(2), 165-179. doi:10.1017/S0272263100011943
- Sheen, R. (1994). A critical analysis of the advocacy of the task-based syllabus. *TESOL Quarterly*, 28(1), 127-151. doi:10.2307/3587202
- Sheen, R. (2003). Focus on form - a myth in the making? *ELT Journal*, 57(3), 225-233. doi:10.1093/elt/57.3.225
- \*Silver, R. E. (1999). *Learning conditions and learning outcomes for second language acquisition: Input, output, and negotiation* (Doctoral dissertation). Retrieved from Dissertations & Theses: A&I. (AAT 9926200)
- Skehan, P. (1998). *A cognitive approach to language learning*. Oxford, UK: Oxford University Press.
- Skehan, P. (2001). Tasks and language performance assessment. In M. S. Bygate (Ed.), *Researching pedagogic tasks: Second language learning, teaching and testing* (pp. 167-185). Harlow, UK: Pearson.
- Skehan, P., & Foster, P. (2001). Cognition and tasks. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 183-205). Cambridge, UK: Cambridge University Press.
- Slavin, R. E. (1986). Best evidence synthesis: An alternative to meta-analysis and traditional reviews. *Educational Researcher*, 15(9), 5-11.

- Spada, N. (1990). Observing classroom behavior and learning outcomes in different second language programs. In J. Richards & D. Nunan (Eds.), *Second language teacher education* (pp. 293-310). Cambridge, UK: Cambridge University Press.
- Spada, N. (1997). Form-focused instruction and second language acquisition: A review of classroom and laboratory research. *Language Teaching*, 30(2), 73-87. doi:10.1017/S0261444800012799
- Spada, N., & Lightbown, P. (2008a). Form-focused instruction: Isolated or integrated? *TESOL Quarterly*, 42(2), 181-208.
- Spada, N., & Lightbown, P. (2008b). Interaction research in second/foreign language classrooms. In A. Mackey & C. Polio (Eds.), *Multiple Perspectives on Interaction* (pp. 157-175). New York: Routledge.
- Spada, N., & Tomita, Y. (2010). Interactions between type of instruction and type of language feature: A meta-analysis. *Language Learning*, 60, 263-308. doi: 10.1111/j.1467-9922.2010.00562.x
- Sternberg, R. J. (Ed.). (1999). *Handbook of creativity*. Cambridge, UK: Cambridge University Press.
- Swain, M. (1985). Communicative competence: Some roles of comprehensible input and comprehensible output in its development. In S. Gass & C. Madden (Eds.), *Input in second language acquisition* (pp. 235-253). Rowley, MA: Newbury House.
- Swain, M. (1991). French immersion and its offshoots: Getting two for one. In B. Freed (Ed.), *Foreign language acquisition: Research and the classroom* (pp. 91-103). Lexington, MA: Heath.
- Swain, M. (1993). The output hypothesis: Just speaking and writing aren't enough. *Canadian Modern Language Review*, 50(1), 158-164.
- Swain, M. (1995). Three functions of output in second language learning. In G. Cook & B. Seidelhofer (Eds.), *Principle and practice in applied linguistics: Studies in honour of H. G. Widdowson* (pp. 125-144). Oxford, UK: Oxford University Press.
- Swain, M. (1998). Focus on form through conscious reflection. In C. Doughty & J. Williams (Eds.), *Focus on form in classroom second language acquisition* (pp. 64-81). Cambridge, UK: Cambridge University Press.
- Swain, M., & Lapkin, S. (1998). Interaction and second language learning: Two adolescent French immersion students working together. *The Modern Language Journal*, 82(3), 320-337. doi:10.2307/329959



- Swain, M., & Lapkin, S. (2001). Focus on form through collaborative dialogue: Exploring task effects. In M. Bygate, P. Skehan, & M. Swain (Eds.), *Researching pedagogic tasks: Second language learning, teaching, and testing* (pp. 99-118). London: Longman.
- Swain, M., & Lapkin, S. (2002). Talking it through: Two French immersion learners' response to reformulation. *International Journal of Educational Research*, 37(3), 285-304. doi:10.1016/S0883-0355(03)00006-5
- Swan, M. (2005). Legislation by hypothesis: The case of task-based instruction. *Applied Linguistics*, 26(3), 376-401. doi:10.1093/applin/ami013
- Tokowicz, N., & MacWhinney, B. (2005). Implicit and explicit measures of sensitivity to violations in second language grammar: An event-related potential investigation. *Studies in Second Language Acquisition*, 27(2), 173-204. doi:10.1017/S0272263105050102
- Tomlinson, B. (2007). Using form-focused discovery approaches. In S. Fotos & H. Nassaji (Eds.), *Form-focused instruction and teacher education*. Oxford, UK: Oxford University Press.
- \*Toth, P. D. (2008). Teacher- and learner-led discourse in task-based grammar instruction: Providing procedural assistance for L2 morphosyntactic development. *Language Learning*, 58(2), 237-283. doi:10.1111/j.1467-9922.2008.00441.x
- Tuz, E. (1993). From controlled practice to communicative activity: Does training transfer? *Temple University Japan Research Studies in TESOL*, 1, 97-108.
- \*Ueno, J. (2005). Grammar instruction and learning style. *Japanese Language and Literature*, 39, 1-25.
- Van den Branden, K. (2006). Introduction: Task-based language teaching in a nutshell. In K. Van den Branden, M. H. Long, & J. C. Richards (Eds.), *Task-based language education: From theory to practice* (pp. 1-16). Cambridge, UK: Cambridge University Press.
- Van den Branden, K. (2007, September). *Task-based language education: From theory to practice... and back again*. Plenary session conducted at the 2<sup>nd</sup> International TBLT Conference. University of Hawaii, Honolulu.
- van Lier, L. (1996). *Interaction in the language curriculum: Awareness, autonomy and authenticity*. New York: Longman.
- VanPatten, B. (1993). Grammar teaching for the acquisition-rich classroom. *Foreign Language Annals*, 26(4), 435-450. doi:10.1111/j.1944-9720.1993.tb01179.x

- VanPatten, B. (1996). *Input processing and grammar instruction in second language acquisition*. Norwood, NJ: Ablex Publishing Corporation.
- VanPatten, B., & Cadierno, T. (1993). Explicit instruction and input processing. *Studies in Second Language Acquisition*, 15(2), 225-243.  
doi:10.1017/S0272263100015394
- VanPatten, B., & Oikkenon, S. (1996). Explanation versus structured input in processing instruction. *Studies in Second Language Acquisition*, 18(4), 495-510.
- Vygotsky, L. S. (1986). *Thought and language* (Kozulin, A., Trans.). Cambridge, MA: MIT.
- Wajnryb, R. (1990). *Grammar dictation*. Oxford, UK: Oxford University Press.
- White, L. (1987). Against comprehensible input: The input hypothesis and the development of second language competence. *Applied Linguistics*, 8(1), 95-100.  
doi:10.1093/applin/8.2.95
- White, L. (1991). Adverb placement in second language acquisition: Some effects of positive and negative evidence in the classroom. *Second Language Research*, 7(2), 133-161. doi:10.1177/026765839100700205
- Widdowson, H. (1978). *Teaching language as communication*. Oxford, UK: Oxford University Press.
- Widdowson, H. (1998). Skills, abilities, and contexts of reality. *Annual Review of Applied Linguistics*, 18, 323-333.
- Widdowson, H. G. (1988). Grammar, nonsense and learning. In W. E. Rutherford & M. S. Smith (Eds.), *Grammar and second language teaching: A book of readings* (pp. 146-155). New York: Newbury House.
- Wigglesworth, G. (2001). Influences on performance in task-based oral assessments. In M. Bygate, P. Skehan, & M. Swain (Eds.), *Task based learning* (pp. 186-209). Harlow, UK: Longman.
- Wildner-Bassett, M. E. (2005). CMC as written conversation: A critical social-constructivist view of multiple identities and cultural positioning in the L2/C2 classroom, *CALICO Journal*, 22(3), 635-656. Retrieved from [https://www.calico.org/html/article\\_160.pdf](https://www.calico.org/html/article_160.pdf)
- Wilkins, D. (1976). *Notional syllabuses*. Oxford, UK: Oxford University Press.
- Williams, J. (1999). Learner-generated attention to form. *Language Learning*, 49(4), 583-625. doi:10.1111/0023-8333.00103

- Williams, J. (2001). The effectiveness of spontaneous attention to form. *System*, 29(3), 325-340. doi:10.1016/S0346-251X(01)00022-7
- Williams, J. (2005). Form-focused instruction. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 671-691). Mahwah, NJ: Lawrence Erlbaum.
- Willis, J. (1996). *A framework for task-based learning*. London: Longman.
- Willis, J. R. (1998, November). Designing and using tasks to promote optimum language development. *The Proceedings of the JALT 24<sup>th</sup> Annual International Conference on Language Teaching/Learning and Educational Materials Expo*. Saitama, Japan: JALT.
- Willis, J. R. (2004). Perspectives on task-based instruction: Understanding our practices, acknowledging different practitioners. In B. L. Leaver & J. R. Willis (Eds.), *Task-based instruction in foreign language education* (pp. 3-44). Washington, DC: Georgetown University Press.
- Willis, D., & Willis, J. (2007). *Doing task-based teaching*. Oxford, UK: Oxford University Press.
- Wong, W. (2005). Input flood. In W. Wong (Ed.), *Input enhancement: From theory and research to the classroom* (pp. 37-47). Boston: McGraw-Hill.
- Yano, Y., Long, M. H., & Ross, S. (1994). The effects of simplified and elaborated texts on foreign language reading comprehension. *Language Learning*, 44(2), 189-219. doi:10.1111/j.1467-1770.1994.tb01100.x
- Zhou, Y.-P. (1991). The effect of explicit instruction on the acquisition of English grammatical structures by Chinese learners. In C. James & P. Garrett (Eds.), *Language Awareness in the classroom* (pp. 254-277). London: Longman.

\*References marked with an asterisk indicate studies included in the meta-analysis.

## Appendixes

## Appendix A

### Abbreviations

## Abbreviations

ANOVA	(One-Way) Analysis of Variance
CI	Confidence Interval
CLT	Communicative Language Teaching
DLIFLC	Defense Language Institute Foreign Language Center
EFL	English as a Foreign Language
ESL	English as a Second Language
ESP	English for Specific Purposes
FFI	Form-Focused Instruction
FL	Foreign Language
FoF	Focus on Form
FoFS	Focus on Forms
FoM	Focus on Meaning
IEP	Intensive English Program
L1	First (i.e., Native) Language
L2	Second Language
LARC	Language Acquisition Research Center
MANOVA	Multivariate Analysis of Variance
NNS	Nonnative Speaker or Nonnative-Speaking
NS	Native Speaker or Native-Speaking
SLA	Second Language Acquisition
TA	Teaching Assistant
TBLT	Task-Based Language Teaching

TESOL	Teachers of English to Speakers of Other Languages
TL	Target Language
TOEFL	Test of English as a Foreign Language
TOEIC	Test of English for International Communication

## Appendix B

### Additional Definitions of Terms



### Additional Definitions of Terms

Accuracy is the extent to which target language output produced by the learner conforms with the target language norms of morphology, syntax, and so forth (R. Ellis, 2003).

Analytic syllabus is an FL or L2 syllabus that is organized in terms of purposes for which people use the language and is built on authentic samples of TL performance necessary to meet these purposes rather than on specific, individual TL items (Long & Robinson, 1998; Wilkins, 1976).

Authentic materials are written or audio passages produced by native speakers of a language for use by other native speakers of this language within the target culture (Brown, 2001) for the purposes of informing, persuading, entertaining them, and so forth (vs. passages created by teachers or course designers specifically for language learners).

Automatization is the process by which declarative knowledge becomes proceduralized through practice and that allows for target language knowledge to be accessed rapidly and effortlessly with minimal demands on the learner's information processing capacity (DeKeyser, 2001).

Clarification request is an interactional strategy used by speakers in order to obtain clarification of the interlocutor's utterance (R. Ellis, 2003), for example, "Excuse me, what do you mean by that"?

Closed task is a task that requires learners to reach a single, correct solution or one of a small finite set of possible solutions (R. Ellis, 2003; Loschky & Bley-Vroman, 1993).

Cognitive complexity is the extent to which cognitive operations involved in completing a task are easy or difficult to execute in terms of the mental processes involved in the execution (Robinson, 2001a).

Communicative competence is the ability to function in the target language that typically is defined as a combination of linguistic, discourse, sociocultural, and strategic competence (Savignon, 2001).

Communicative Language Teaching (CLT) is an approach to teaching that is directed at developing ability to communicate in the language and perform a wide range of functions that native speakers of the target language normally perform within the target culture (Canale & Swain, 1980).

Comprehensible input is the authentic target language input that is at a level slightly beyond the learners' current competence level that can still be comprehended by them (Krashen, 1985), perhaps through interactions with native speakers or peers that involve negotiation of meaning or through elaboration of the input (Long, 1983).

Comprehension check is an interactional strategy used by speakers to check whether their preceding utterance has been understood by the interlocutor (R. Ellis, 2003), for example, "Do you know what I mean"?

Confirmation check is an interactional strategy used by speakers to make sure that they have understood correctly what the interlocutor has said (R. Ellis, 2003), for example, "You said you were not going to the party, right"?

Consciousness-raising activity is an activity that engages learners in thinking and communicating about target language or its specific features, rather than about real-world information, with a purpose of raising their understanding of the functioning of these language features (Fotos, 1994).

Controlled processing is processing that occurs when learners utilize conscious effort and attention to their own performance in the target language, involves declarative

knowledge, and is demanding of the learners' information-processing capacity (R. Ellis, 2003).

Convergent task is a task that requires the participants to agree to a common solution or task outcome (Duff, 1986; R. Ellis, 2003).

Counterbalancing refers to test design in which the order of presentation of test items or tasks is different for different participants in order to prevent the so-called test learning effects (Mackey & Gass, 2005).

Custom-made test is a test that is tailored to individual learners because it is based on the errors (e.g., in grammatical structures) made by these learners on a pretest in an attempt to measure the specific effect of the instructional treatment on each individual learner with his or her individual state of interlanguage development (Mackey & Goo, 2007).

Declarative knowledge is knowledge about the target language (e.g., a grammar rule) that has not yet been proceduralized and automatized (R. Ellis, 2003).

Dictogloss is an activity that requires learners to reconstruct a short text that they have heard presented at normal rate of speech as close to the original as possible (Wajnryb, 1990).

Discourse is spoken or written language (Brown, 2001).

Display question is a question to which the speaker already knows the answer intended to elicit a display of target language use rather than providing information (Long, 1997).

Divergent task is a task that does not require the participants to produce a common solution or outcome but rather encourages them to pursue differing agendas or defend opposing views (Duff, 1986; R. Ellis, 2003).

Explicit instructional technique is a classroom technique that involves conscious

cognitive processing by directing the learners' attention overtly to language features that they need to learn (Norris & Ortega, 2000).

Explicit linguistic knowledge is verbalizable knowledge about the target language, for example, knowledge of a particular grammar rule (R. Ellis, 2003).

Fluency is the extent to which the target language output produced by the learner approximates the normal rate of delivery and is free of hesitation pauses, reformulations caused by lack of linguistic competence, and so forth (Doughty & Long, 2006).

Form-Focused Instruction (FFI) is any planned or incidental activity that focuses the learners' attention on language form, regardless of the nature of this activity (i.e., the activity can represent either Focus on Form or Focus on Forms; R. Ellis, 2001).

Fossilization is a phenomenon characterized by persistent retention of ungrammatical language forms in the learner's interlanguage despite the presence of opportunities to improve (Long, 2006; Selinker, 1972).

Implicit instructional technique is a technique that promotes TL learning that takes place without the learner's awareness while the learner is engaged in meaning-based activities without overt attention to form (R. Ellis, 2003; Norris & Ortega, 2000).

Implicit linguistic knowledge is intuitive knowledge that the learner may not be able to verbalize that manifests itself in the ability to communicate fluently in the target language or to evaluate whether a target language string is formulated appropriately (i.e., whether it adheres to target language norms; R. Ellis, 2003).

Implicit techniques are error correction and other instructional techniques that teachers use to draw the learner's attention indirectly to linguistic form without interrupting the flow of meaningful communication (R. Ellis, 2003).

Information-gap task is a task in which one participant holds information that the other participant(s) do(es) not have, and the participants must exchange information in order for the task to be completed successfully (R. Ellis, 2003; Prabhu, 1987).

Input hypothesis is a hypothesis formulated by Krashen (1985) that posits that target language acquisition occurs as a result of comprehending input slightly above the learners' current level.

Input-processing instruction is a subset of instructional techniques that are specifically aimed at getting learners to process the form-meaning connections associated with a specific linguistic feature before they are asked to produce their own output containing this feature (Lee & VanPatten, 2003; VanPatten, 1996).

Intake is the subset of the input that has been noticed and processed by the learner (Schmidt, 1990).

Interaction hypothesis is a hypothesis formulated by Long (1981, 1996), first presented in 1980, that posits that learners acquire target language as a result of attending to linguistic features in the process of negotiating for meaning while trying to overcome miscommunication.

Interface position is the position that claims that explicit linguistic knowledge can be converted to implicit knowledge through practice of specific target language features (R. Ellis, 2003).

Interlanguage is the representation of the target language in the mind of the learner, the idiosyncratic implicit linguistic system that the learner has built at a specific stage of language development (Long, 2006; Selinker, 1972).

Jigsaw task is a task where the input material is divided between the participants so that they all are required to exchange information in order to complete the task successfully (i.e., a two-way information-gap task; R. Ellis, 2003; Pica, Kanagy, & Falodun, 1993).

Language aptitude is a special subset of abilities involved in learning a second or foreign language (Skehan, 1998).

Lexis is all the words, or vocabulary items, in a language (Doughty & Long, 2006).

Metalinguistic refers to processes that involve thinking or talking about the target language (Doughty & Long, 2006).

Modified output is the process that occurs when a participant in a conversation reformulates the original utterance as a reaction to feedback received from the interlocutor, for example, when the interlocutor signals lack of comprehension (R. Ellis, 2003; Long, 1996).

Morphology is a branch of linguistics that studies word forms resulting from grammatical rules governing the language (e.g., noun declension, verb conjugation, etc.; Doughty & Long, 1996).

Negotiation of form is the process by which two or more interlocutors try to resolve a linguistic problem that resulted from inappropriate use of a specific language item (Long, 1996).

Negotiation of meaning is the process by which two or more interlocutors try to resolve a communication problem that has been caused by lack of comprehension of intended meaning (Foster, 1998; Long, 1996).

Noninterface position is the position that claims that explicit linguistic knowledge does not get converted into implicit knowledge necessary for fluent communication in the target language (R. Ellis, 2003).

Noticing is a cognitive process that involves attending to linguistic form in the input that the learners receive or the output they produce (Schmidt, 1993).

One-way task is an information-gap task where only one of the participants has to communicate information to the other(s) who do(es) not hold any information that needs to be communicated for successful completion of the task (R. Ellis, 2003).

Open task is a task that does not have a predetermined solution, and, therefore, many outcomes are acceptable (R. Ellis, 2003; Loschky & Bley-Vroman, 1993).

Opinion-gap task is a task that requires the participants to exchange opinions on an issue (R. Ellis, 2003; Prabhu, 1987).

Output hypothesis is a hypothesis formulated by Swain (1985) that posits that learner-produced output is necessary for target language acquisition in addition to input, and that acquisition is facilitated when learners are pushed to produce target language output that is accurate and precise (Swain, 1993).

Pedagogic task is a task that is designed to elicit communicative target language use that, unlike a in a real-world task, does not resemble a real-world event or function but, nevertheless, leads to patterns of language use similar to those found in the real world (e.g., the spot-the-difference picture task; R. Ellis, 2003).

Pragmatics is a branch of linguistics that studies norms of appropriateness in social interaction and the ways in which context contributes to the meaning of utterances

(Brown, 2001). For example, “It is cold in here” may mean “Close the window” or “We can store food here” depending upon the circumstances of the interaction.

Pretask planning is the process by which learners plan what they are going to do and say during task performance before the task commences (R. Ellis, 2003; Foster & Skehan, 1996).

Procedural knowledge is knowledge that is automatized and, therefore, can be accessed rapidly and relatively effortlessly during task performance (R. Ellis, 2003).

Productive language skills are skills that involve production of target language output by the learner, that is, speaking and writing, as opposed to receptive skills that only involve comprehension (Brown, 2001).

Proficiency test is a foreign or second language test that aims to assess global competence in the target language and is not limited to any specific language items, curriculum, or course. A typical example of a standardized proficiency test is the Test of English as a Foreign Language (TOEFL; Brown, 2001).

Pushed output is output that is created when learners are pushed to produce in the target language, especially when they are pushed to produce accurately and concisely (Keck et al., 2006; Swain, 1993).

Reasoning-gap task is a task that encourages the participants to engage in reasoning or figuring out a solution to a problem collaboratively while interacting in the target language (R. Ellis, 2003; Prabhu, 1987).

Recast is an utterance produced by the teacher or a peer that rephrases the learner’s preceding utterance in a more appropriate, native-like manner without changing its meaning (Lyster & Ranta, 1997).



Receptive language skills are skills that require comprehension, but not production, on behalf of the learner, that is, reading and listening (Brown, 2001).

Scaffolding is a subset of instructional techniques that can be used to help the learner accomplish a task successfully, typically through helpful interaction with more proficient partners (R. Ellis, 2003).

Semantics is a branch of linguistics that studies the ways in which words and word-forms of a language convey meaning (Brown, 2001).

Structure-based production task is a focused task designed with a goal of eliciting production of a specific structure (Loschky & Bley-Vroman, 1993).

Syntax is a branch of linguistics that studies rules of arranging words into grammatically appropriate clauses and sentences (Doughty & Long, 2006).

Synthetic syllabus is an FL or L2 syllabus that is based on gradual accumulation of language items that are taught separately and step by step (Wilkins, 1976). The most common example of a synthetic syllabus is the so-called structural syllabus that is based on teaching grammatical structures one at a time in a linear fashion (Long & Robinson, 1998).

Task cycle is a lesson design that consists of three stages: pretask, during task, and posttask (R. Ellis, 2003).

Text-reconstruction task is an activity that requires learners to reconstruct all or only the missing parts of a passage that they previously read or heard, frequently in order to elicit use of specific structures seeded in the passage (R. Ellis, 2003).

Transfer-appropriate processing is the type of processing that is said to take place when the initial encoding of information happens under the same conditions under which this

information will be retrieved later (DeKeyser, 2007; Lightbown, 2007). In other words, the degree of success in retrieving information encoded in memory is determined, among other factors, by the relationship between how this information was encoded initially and how it is retrieved later (i.e., retrieval will be most successful when the processes that are involved in encoding are the same processes that are active during retrieval). For example, filling in the blanks with correct grammatical endings does not represent transfer-appropriate processing if the learner's goal is using grammar correctly in communication.

Two-way task is an information-gap task where the information to be exchanged is split between two or more participants (R. Ellis, 2003).

Uptake is the part of the processed input, or intake, that has been internalized and is now available for subsequent use by the learner (R. Ellis, 2003).

Washback is the effect that a test has on teaching practices (Bailey, 1996) such as when teachers focus their classroom instruction on specific types of tasks because they are included in the test.

## Appendix C

### Coding Form

## Coding Form

Coder: \_\_\_\_\_

Date: \_\_\_\_\_

## Identification of Studies

1. Study ID number

\_\_\_\_\_

2. Author name(s)

\_\_\_\_\_

3. Year of publication

\_\_\_\_\_ / Unknown

4. Source (provide APA citation)

\_\_\_\_\_

## Outcome Features

1. Construct measured (e.g., acquisition of target structure X)

\_\_\_\_\_

2. Source of above construct definition (circle one)

a. Defined by primary researcher(s)

b. Inferred by rater

3. Type of outcome (circle all that apply)

a. Posttest scores

b. Gain scores

c. Both

## 4. Pretest

Present \_\_\_\_\_

Not present \_\_\_\_\_

4.1. If pretest is present, specify type of test (circle all that apply)

- a. Metalinguistic grammaticality judgment
  - b. Selected response
  - c. Constrained-constructed response
  - d. Free-constructed response (specify prompt)
- 

- e. Oral communication task (specify task)

---

4.2. If pretest is present, specify whether test counterbalancing measures are used

- a. Yes (specify counterbalancing measures)
- 

- b. No

4.3. If pretest is present, specify test congruency with TBLT methodology

- a. Congruent
- b. Not congruent

## 5. Immediate posttest

Present \_\_\_\_\_

Not present \_\_\_\_\_

5.1. If an immediate posttest is present, specify type of test (circle all that apply)

- a. Metalinguistic grammaticality judgment
- b. Selected response
- c. Constrained-constructed response

- d. Free-constructed response (specify prompt)
- 

- e. Oral communication task (specify task)
- 

- 5.2. If an immediate posttest is present, specify whether test counterbalancing measures are used

- a. Yes (specify counterbalancing measures)
- 

- b. No

- 5.3. If an immediate posttest is present, specify test congruency with TBLT methodology

- a. Congruent  
b. Not congruent

6. Delayed posttest

Present \_\_\_\_\_

Not present \_\_\_\_\_

- 6.1. If a delayed posttest is present, specify type of test (circle all that apply):

- a. Metalinguistic grammaticality judgment  
b. Selected response  
c. Constrained-constructed response  
d. Free-constructed response (specify prompt)
- 

- e. Oral communication task (specify task)
-

6.2. If a delayed posttest is present, specify whether test counterbalancing measures are used

a. Yes (specify counterbalancing measures)

---

b. No

6.3. If a delayed posttest is present, specify test congruency with TBLT methodology

a. Congruent

b. Not congruent

6.4. If a delayed posttest is present, specify length of delay in days \_\_\_\_\_

6.5. If (an)other delayed posttest(s) is or are present, specify length of delay in days and other relevant information here

---



---

### Methodological Features

1. Type of report source (circle one)

a. Peer-reviewed journal

b. Not peer-reviewed journal

c. Doctoral dissertation

d. Master thesis

e. Book chapter

f. Conference report

g. Other unpublished report (specify) \_\_\_\_\_

## 2. Educational setting (circle one)

- a. High school
  - b. Undergraduate level
  - c. Graduate level
  - d. IEP (Intensive English Program)
  - e. ESP program (English for Specific Purposes)
  - f. Adult education
  - g. Other (specify)
- 

- h. Unknown

## 3. Control and comparison groups (circle all that apply)

- a. One control group

Specify number of participants in the control group \_\_\_\_\_

- b. One comparison group

Specify number of participants in the comparison group \_\_\_\_\_

- c. More than one control group (specify number of groups) \_\_\_\_\_

Label all control groups and specify number of participants in each

---



---

- d. More than one comparison group (specify number of groups) \_\_\_\_\_

Label all comparison groups and specify number of participants in each

---



---

- e. No control and no comparison groups



4. Experimental (task-based interaction treatment) groups (circle one that applies)

- a. One experimental group

Specify number of participants in the experimental group \_\_\_\_\_

- b. More than one experimental group (specify number of groups) \_\_\_\_\_

Label all control groups and specify number of participants in each

---



---

5. Basis for determining participant TL proficiency level

- a. Impressionistic judgment

- b. Institutional placement test

- c. Institutional course enrollment

- d. Standardized test (specify) \_\_\_\_\_

- e. Other (specify) \_\_\_\_\_

- f. Unknown

6. Presence of pretest (transfer from 4 under Outcome Features)

Present \_\_\_\_\_

Not present \_\_\_\_\_

- 6.1. If a pretest is present, specify whether participants were eliminated on the basis of the pretest

- a. Yes (specify reasons)

---

- b. No

7. Target language

- 7.1. Target language (TL; circle one)

- a. English
- b. Other than English (specify language) \_\_\_\_\_

7.2. If other than English, specify language group (MacWhinney, 1995; circle one)

- a. Language group I
- b. Language group II
- c. Language group III
- d. Language group IV
- e. Language group V

7.3. Specify language learning setting

- a. Foreign language (FL)
- b. Second language (L2)

8. Outcome measure (circle one)

- a. Standardized test
- b. Uniform researcher-made test
- c. Custom-designed researcher-made test
- d. Uniform teacher-made test
- e. Custom-designed teacher-made test
- f. Other (specify)\_\_\_\_\_
- g. Unknown

9. Statistics reported (circle all that apply)

- a. Means/ Standard deviations (specify) \_\_\_\_\_
- b. *t* test/ Degrees of freedom (specify) \_\_\_\_\_

- c.  $F$  test/ Degrees of freedom (specify) \_\_\_\_\_
- d.  $p$  level/ Sample size (specify) \_\_\_\_\_
- e. Proportion of participants who experienced gain \_\_\_\_\_
- f. Effect size (specify in 10)

10. Effect size \_\_\_\_\_

10.1. Type of effect size value (circle one)

- a. Standardized mean difference
- b. Standardized mean gain

10.2. Source of effect size value (circle one)

- a. Reported
- b. Calculated
- c. Estimated from probability levels

11. Treatment duration (circle one and specify)

11.1. Actual length of treatment (combined if several sessions)

- a. \_\_\_\_\_ minutes
- b. \_\_\_\_\_ hours

11.2. Does this include pretask and posttask phases?

- a. Yes
- b. No
- c. Unknown

11.3. Specify number of sessions (per individual participant) \_\_\_\_\_

11.4. Treatment delivered over the course of

- a. \_\_\_\_\_ weeks
- b. \_\_\_\_\_ months

c. \_\_\_\_\_ semesters

12. Instructor equivalence between treatment group and control or comparison group

(circle one)

- a. Same instructor
- b. Different instructor
- c. Unknown
- d. Not applicable (e.g., if gain scores for one group are reported)

13. Student equivalence (circle one)

- a. Random
- b. Statistical control
- c. Students self-select
- d. Intact class
- e. Unknown
- f. Not applicable (e.g., if gain scores for one group are reported)

Learner Characteristics: All Groups

1. Number of learners \_\_\_\_\_ / Unknown

2. Gender

Number of males \_\_\_\_ / Unknown      Number of females \_\_\_\_ / Unknown

3. Average age \_\_\_\_\_ / Unknown

4. Age range \_\_\_\_\_ / Unknown

5. L1 (circle one)

- a. Same L1 for all learners (specify L1) \_\_\_\_\_
- b. Learners have different L1s (specify L1s) \_\_\_\_\_

c. Unknown

6. Degree of L1-L2 similarity (if applicable) \_\_\_\_\_

7. TL proficiency level (circle one)

a. Low beginner

b. Beginner

c. High beginner

d. Low intermediate

e. Intermediate

f. High intermediate

g. Advanced

h. Mixed (specify) \_\_\_\_\_

i. Unknown

Additional information \_\_\_\_\_

Learner Characteristics: Treatment Group(s)

8. Number of learners \_\_\_\_\_ / Unknown

9. Gender

Number of males \_\_\_\_\_ / Unknown

Number of females \_\_\_\_\_ / Unknown

10. Average age \_\_\_\_\_ / Unknown

11. Age range \_\_\_\_\_ / Unknown

12. L1 (circle one)

a. Same L1 for all learners (specify L1) \_\_\_\_\_

b. Learners have different L1s (specify L1s) \_\_\_\_\_

c. Unknown

13. Degree of L1-L2 similarity (if applicable) \_\_\_\_\_

14. TL proficiency level (circle one)

- a. Low beginner
- b. Beginner
- c. High beginner
- d. Low intermediate
- e. Intermediate
- f. High intermediate
- g. Advanced
- h. Mixed (specify) \_\_\_\_\_
- i. Unknown

15. Additional information \_\_\_\_\_

Learner Characteristics: Control Group

Present \_\_\_\_\_ Not present \_\_\_\_\_

1. Number of learners \_\_\_\_\_ / Unknown

2. Gender

Number of males \_\_\_\_ / Unknown      Number of females \_\_\_\_ / Unknown

3. Average age \_\_\_\_\_ / Unknown

4. Age range \_\_\_\_\_ / Unknown

5. L1 (circle one)

- a. Same L1 for all learners (specify L1) \_\_\_\_\_
- b. Learners have different L1s (specify L1s) \_\_\_\_\_
- c. Unknown

6. Degree of L1-L2 similarity (if applicable) \_\_\_\_\_

7. TL proficiency level (circle one)

a. Low beginner

b. Beginner

c. High beginner

d. Low intermediate

e. Intermediate

f. High intermediate

g. Advanced

h. Mixed (specify) \_\_\_\_\_

i. Unknown

8. Additional information \_\_\_\_\_

Learner Characteristics: Comparison Group(s) (specify type) \_\_\_\_\_

Present \_\_\_\_\_

Not present \_\_\_\_\_

1. Number of learners \_\_\_\_\_ / Unknown

2. Gender

Number of males \_\_\_\_\_ / Unknown

Number of females \_\_\_\_\_ / Unknown

3. Average age \_\_\_\_\_ / Unknown

4. Age range \_\_\_\_\_ / Unknown

5. L1 (circle one)

a. Same L1 for all learners (specify L1) \_\_\_\_\_

b. Learners have different L1s (specify L1s) \_\_\_\_\_

c. Unknown

6. Degree of L1-L2 similarity (if applicable) \_\_\_\_\_

7. TL proficiency level (circle one)

- a. Low beginner
- b. Beginner
- c. High beginner
- d. Low intermediate
- e. Intermediate
- f. High intermediate
- g. Advanced
- h. Mixed (specify) \_\_\_\_\_
- i. Unknown

8. Additional information \_\_\_\_\_

Treatment design and pedagogical features

Specify task \_\_\_\_\_

If multiple tasks are used in this treatment, duplicate this part of the Coding Form and fill out for each task. Specify the total number of the tasks used in the treatment here \_\_\_\_\_

1. Source of task (circle one)

- a. Designed by teacher
- b. Designed by researcher
- c. Designed by curriculum developer
- d. Other (specify)
- e. Unknown



## 2. Task type

### 2.1. Task design (circle one)

- a. Information-gap
  - b. Jigsaw
  - c. Problem-solving
  - d. Decision-making
  - e. Opinion-gap
  - f. Information transfer
  - g. Role-play
  - h. Narrative
  - i. Compound (specify, e.g., information-gap and decision-making)
- 

- j. Other
- k. Unknown

### 2.2. Information flow (circle one)

- a. One-way
- b. Two-Way
- c. Unknown

### 2.3. Intended outcome (circle one)

- a. Closed
- b. Open
- c. Unknown

2.4. Participants' goals (circle one)

- a. Convergent
- b. Divergent
- c. Unknown

3. Pretask stage

3.1. Conducted by (circle one)

- a. Teacher
- b. Researcher
- c. Other (specify) \_\_\_\_\_

3.2. Components (circle all that apply)

- a. Rule review
- b. Modeling of target structure
- c. Exercises with focus on target structure
- d. Learner opportunity for pretask planning
- e. Other (specify) \_\_\_\_\_
- f. Unknown

4. During-task stage

4.1. Task set up by (circle one)

- a. Teacher
- b. Researcher
- c. Other (specify) \_\_\_\_\_

## 4.2. Interaction type (circle one)

## a. Learner-to-learner (if applicable, circle one)

- i. Students receive linguistic help from teacher, researcher, or other NS
- ii. Students receive strategy help from teacher, researcher, or other NS
- iii. Students receive both linguistic and strategy help
- iv. Students receive no linguistic and no strategy help
- v. Unknown

## b. NS-to-learner (if applicable, circle one)

- i. Teacher-led
- ii. Researcher-led
- iii. TA-led
- iv. Other NS-led

## 4.3. Error correction (circle one)

- a. Provided
- b. Not provided
- c. Unknown

## 5. Posttask stage (circle all that apply)

## a. Feedback on errors (if present, circle all types of feedback that apply)

- i. Oral feedback
- ii. Written feedback
- iii. Other (specify) \_\_\_\_\_

## iv. Unknown

- b. Rule review
- c. Exercises with focus on target structure
- d. Other (specify) \_\_\_\_\_
- e. Unknown

## 6. Target structure (specify) \_\_\_\_\_

If multiple grammatical structures are targeted by the same treatment, duplicate this part of the Coding Form and fill out for each structure. Specify the number of the target structures here \_\_\_\_\_

## 6.1. Type (circle one)

- a. Morphological
- b. Syntactic
- c. Morphosyntactic
- d. Unknown

## 6.2. Complexity (circle one)

- a. Simple
- b. Complex
- c. Unknown

## 6.3. Ambiguity (circle one)

- a. Ambiguous
- b. Unambiguous
- c. Unknown

6.4. Degree of task-essentialness (circle one)

- a. Task-natural
- b. Task-useful
- c. Task-essential

6.5. Determination of task-essentialness (circle one)

- a. Reported in the study
- b. Inferred by rater

6.6. Evidence of target structure use during task completion (circle all that apply)

- a. Not available
- b. Interaction transcripts available
- c. Usage counts available
- d. Other available (specify) \_\_\_\_\_

7. Learner attitudes toward TBLT (circle one)

- a. Favorable
- b. Unfavorable
- c. Unknown

8. Teacher/ TA attitudes toward TBLT (circle one)

- a. Favorable
- b. Unfavorable
- c. Unknown
- d. Not applicable (e.g., treatment conducted by the researcher)

9. Teacher/ TA familiarity with TBLT (circle all that apply)

- a. Training provided before treatment
- b. Received training previously

- c. Used TBLT previously
- d. Unknown
- e. Not applicable (e.g., treatment conducted by the researcher)

10. Additional information \_\_\_\_\_

#### Quality of Study

1. Publication bias/ Review process (circle one)

- a. Peer-reviewed
- b. Not peer-reviewed
- c. Unknown

2. Attrition for control group (circle one)

- a. Known (specify) \_\_\_\_\_(%)
- b. Unknown

3. Attrition for comparison group (circle one)

- a. Known (specify) \_\_\_\_\_(%)
- b. Unknown

4. Attrition for treatment group (circle one)

- a. Known (specify) \_\_\_\_\_(%)
- b. Unknown

5. Validity of outcome measure(s) (circle one)

- a. Information reported (specify) \_\_\_\_\_
- b. Not reported

6. Reliability of outcome measure(s) (circle one)

- a. Information reported (specify) \_\_\_\_\_
- b. Not reported

## Appendix D

### Draft Electronic Message Requesting a Copy of Study Report

Draft Electronic Message Requesting a Copy of Study Report

Dear Professor,

I am a doctoral candidate at the University of San Francisco, School of Education, Department of Learning and Instruction. My research field is Second Language Acquisition. I am conducting a meta-analysis of the effectiveness of task-based interaction in form-focused instruction of adult learners in foreign and second language teaching. It appears that the research study you have conducted may be a candidate for inclusion in my meta-analysis. I will be very appreciative if you kindly forward me a copy of your study report/dissertation/thesis. [The Interlibrary Loan Department at the USF library has informed that the only available copy of your dissertation/thesis is held at the X University as a non-circulating item.]

Marina Cobb  
Doctoral Candidate  
Learning and Instruction  
University of San Francisco